UNIVERSITÉ NICE SOPHIA ANTIPOLIS — UFR Sciences

École Doctorale Sciences Fondamentales et Appliquées

THÈSE

pour obtenir le grade de

Docteur en Sciences de l'Université Nice Sophia Antipolis

Discipline : Mathématiques

présentée et soutenue par

Mathieu SART

Estimation par tests

Thèse encadrée par : M. Yannick BARAUD

Rapporteurs : M. Olivier CATONI
Mme. Sara VAN DE GEER

Soutenue le 25 novembre 2013 devant le jury composé de :

M.	Yannick BARAUD	Université Nice Sophia Antipolis	Directeur de thèse
M.	Lucien BIRGÉ	Université Pierre et Marie Curie	Examinateur
M.	Olivier CATONI	ENS, Paris	Rapporteur
M.	Matthieu LERASLE	Université Nice Sophia Antipolis	Examinateur
M.	Pascal MASSART	Université Paris-Sud	Examinateur
M.	Vladimir SPOKOINY	WIAS, Berlin	Examinateur

Laboratoire Jean-Alexandre Dieudonné, Parc Valrose, 06108 Nice Cedex 2

Remerciements

Mes premiers remerciements vont à Yannick sans qui cette thèse n'aurait pu voir le jour. Merci pour la confiance et le soutien que vous m'avez accordé durant ces années. J'ai beaucoup appris à vos côtés, et je tenais à vous témoigner, par ces quelques lignes, toute ma gratitude pour avoir encadré cette thèse.

Je suis très honoré que Sara van de Geer et Olivier Catoni aient accepté de rapporter cette thèse. Je suis très reconnaissant à Lucien Birgé, Matthieu Lerasle, Pascal Massart et Vladimir Spokoiny d'avoir accepté de faire partie de mon jury de soutenance.

Merci Lucien pour votre gentillesse et pour la bienveillance avec laquelle vous avez considéré mon travail. Ce serait peu de dire que vos travaux ont influencé ma vision des statistiques !

Mon apprentissage des statistiques a véritablement commencé au M2 de l'université Paris-sud, et je tiens à remercier les professeurs que j'ai eu pour les cours passionnants que j'y ai reçu.

Ce fut un plaisir de travailler au laboratoire Dieudonné et mes plus sincères remerciements vont à Elisabeth Gassiat sans qui je ne serai jamais venu à Nice.

J'ai une pensée particulière pour l'équipe probabilités et statistiques, et je souhaite remercier l'ensemble de ses membres pour l'accueil qui m' a été fait. Merci Thomas pour ta bonne humeur et ta sympathie tout au long de ces années. Merci Christine pour pour ton amitié, ton sens de l'écoute et de ton aide. Patricia, je voulais te dire que ce fut aussi un réel plaisir de discuter avec toi ces années. Matthieu, j'espère bien que l'on sera amené à travailler ensemble par la suite, et je voulais te remercier en particulier pour l'intérêt que tu portes à mon travail.

Le monitorat constitue une part importante de travail et cette tâche n'aurait pas été si agréable sans tous ceux avec lesquels j'ai collaboré.

Merci à Hélène Politano pour son aide pour tout ce qui concerne l'école doctorale. Je fais une spéciale dédicace à Philippe Maisonobe pour sa bonne humeur constante et tous ces moments partagés.

La thèse n'aurait pas été la même sans vous, les doctorants, nouveaux et anciens et je voulais vous dire à quel point j'ai apprécié les moments que nous avons passé ensemble, que ce soit au labo ou à l'extérieur. Je me rappelle de ces discussions philosophiques à n'en plus finir, de ces parties de baseball, de ces magnifiques paysages lors de nos randonnées dans les montagnes du Mercantour. Je ne peux pas ne pas mentionner ces moments inoubliables lors de nos sorties au

ski que ce soit à Isola, Auron, ou Valberg. J'aimerais vous dire à tous un grand merci.

Merci à ma famille pour leur soutien sans faille. Mes sincères félicitations à Audrey et François.

Je termine ces remerciements par celle qui se reconnaîtra, celle qui a le courage de me supporter et de m'attendre pendant tout ce temps.

Table des matières

CHAPITRE 1

Introduction

L'objet de cette thèse est la construction d'estimateurs à partir de tests et l'étude de leurs propriétés. Elle est constituée de quatre chapitres pouvant être lus indépendamment les uns des autres.

Nous présentons ces techniques d'estimation dans l'introduction. Nous construisons ensuite des estimateurs de ce type pour démontrer de nouveaux résultats de sélection de modèles dans deux cadres statistiques différents : celui de l'estimation des intensités de processus de Poisson avec covariables (Chapitre 2) et celui de l'estimation de la densité de transition d'une chaîne de Markov (Chapitre 3). Nous nous intéressons également à la pratique en proposant dans les Chapitres 3 et 4 de nouvelles procédures pour construire plus rapidement des estimateurs de ce type. Plus précisément, nous étudions une nouvelle règle de sélection d'estimateurs constants par morceaux dans le Chapitre 3 et une nouvelle procédure dédiée à l'estimation paramétrique et robuste d'une densité dans le Chapitre 4.

Le plan de l'introduction est le suivant. Nous commençons dans la première section par illustrer l'intérêt des procédures basées sur des tests en donnant un petit aperçu des résultats qu'elles permettent de démontrer dans le cadre de l'estimation d'une densité. Nous expliquons plus en détail les résultats que l'on peut démontrer avec ces procédures dans la Section 2. Les procédures sont décrites dans la Section 3 et nous discutons de leur implémentation pratique dans la Section 4. Les résultats, ainsi que les procédures que nous présentons dans ces quatre sections sont principalement ceux de l'article clé de Birgé (2006). Nous exposons dans la Section 5 les différents problèmes d'estimation qui ont été étudiés dans la littérature à l'aide de tests. La dernière section est consacrée à une présentation des chapitres.

1. Quelques résultats illustratifs

1.1. Cadre statistique de référence. Nous nous plaçons pour simplifier dans le cadre de l'estimation d'une densité. Nous observons n variables aléatoires $X_1, \ldots, X_n$, indépendantes et identiquement distribuées définies sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. Nous supposons que chaque variable aléatoire X_i admet une densité s inconnue par rapport à la

mesure de Lebesgue et nous souhaitons l'estimer à l'aide des observations $X_1, \dots, X_n$.

Nous considérons l'ensemble $\mathbb{L}^1_{\text{dens}}(\mathbb{R}^d)$ des densités sur $\mathbb{R}^d$ et nous le munissons de la distance de Hellinger h définie par

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}^d} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}x \quad \text{pour toutes densités } f, g \in \mathbb{L}^1_{\text{dens}}(\mathbb{R}^d).$$

Nous pouvons modéliser l'information que nous avons sur s en considérant un sous-ensemble $\mathscr{F}$ de $\mathbb{L}^1_{\text{dens}}(\mathbb{R}^d)$. L'idée est que s devrait appartenir à $\mathscr{F}$, ou plus raisonnablement, devrait être "proche" de $\mathscr{F}$, ce qui signifie que $\inf_{f\in\mathscr{F}} h^2(s,f)$ est "petit". Le but est alors de construire un estimateur $\hat{s}$ à valeurs dans ce modèle $\mathscr{F}$, "proche" de s au sens où $\mathbb{E}\left[h^2(s,\hat{s})\right]$ est "petit". Typiquement, $\mathscr{F}$ peut être un ensemble de densités indexé par un sous-ensemble Θ de $\mathbb{R}^d$. Dans ce cas, on dit que $\mathscr{F}$ est un modèle paramétrique. Dans le cas contraire, le modèle est dit non paramétrique.

Dans toute la thèse, $C, C', C'', \dots$ sont des constantes qui pourront varier de ligne en ligne.

1.2. Deux exemples en estimation paramétrique. En estimation paramétrique, les estimateurs construits à partir de tests offrent une alternative à l'estimateur du maximum de vraisemblance. Donnons deux exemples.

1.2.1. Modèle de loi uniforme. Considérons le cas où la loi des X_i est suspectée appartenir à l'ensemble des lois uniformes sur l'intervalle $[0,\theta]$. Cela correspond à considérer le modèle paramétrique $\mathscr{F} = \{f_\theta,\, \theta > 0\}$ où $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$ et à supposer que s n'est pas trop loin de $\mathscr{F}$. La Proposition 1 de Birgé (2006) montre que l'on peut construire l'estimateur $\hat{s}$ de la proposition suivante.

Proposition 1. *Il existe un estimateur $\hat{s}$ de s de la forme $\hat{s} = f_{\hat{\theta}}$ tel que*

$$C\mathbb{E}\left[h^2(s,\hat{s})\right] \le \inf_{\theta>0}\left\{h^2(s,f_\theta) + \frac{\max\left\{\log\left(|\log\theta|/\Gamma_n\right),1\right\}}{n}\right\}$$

où $C > 0$ est une constante universelle et où

$$\Gamma_n = 33.6\times 10^5 n^{-1}\left(4.5\exp\left[\max\left\{(n/84),2\right\}\right]-1\right).$$

Lorsque la distribution des données est effectivement uniforme sur un intervalle de la forme $[0,\theta]$, ce résultat montre que le risque de l'estimateur est au plus $C^{-1}\max\left\{\log\left(|\log\theta|/\Gamma_n\right),1\right\}/n$, ce qui est égal à C^{-1}/n à moins que $|\log\theta|$ soit très grand. Il est intéressant de comparer cette borne de risque au risque de l'estimateur du maximum de vraisemblance $\hat{\theta}_{\text{mle}} = \max_{1\le i\le n} X_i$: $\mathbb{E}[h^2(s,f_{\hat{\theta}_{\text{mle}}})] = (2n+1)^{-1}$. Lorsque le modèle est correct, le risque des deux estimateurs décroit à la même vitesse n^{-1}. En revanche, la constante C^{-1} fournit par la théorie est malheureusement grande, ce qui signifie que le risque de l'estimateur $\hat{s}$ peut être supérieur au risque du maximum de vraisemblance. Nous verrons cependant dans les simulations (page 119), qu'un estimateur construit par tests peut avoir un risque légèrement inférieur à celui du maximum de vraisemblance lorsque ce modèle est vraie.

La raison pour laquelle nous acceptons de perdre un peu sur le risque est que nous ne supposons pas que la vraie densité sous-jacente s appartient au modèle $\mathscr{F}$. Ce modèle est utilisé comme étant une classe de fonctions permettant d'approcher s. En particulier, le risque de l'estimateur reste raisonnable lorsque le modèle est légèrement incorrect, ce qui montre que l'estimateur est robuste. Par exemple, si la vraie distribution des données s est un mélange de deux lois uniformes

$$s = 10\left[(1-2n^{-1})\mathbb{1}_{[0,1/10]} + 2n^{-1}\mathbb{1}_{[9/10,1]}\right]$$

alors $h^2(s, f_{1/10}) = \mathcal{O}(n^{-1})$, et on peut déduire de la proposition que le risque de $\hat{s}$ reste majoré par $\mathbb{E}[h^2(s,\hat{s})] = \mathcal{O}(n^{-1})$. Cette propriété de robustesse n'est pas partagée par l'estimateur du maximum de vraisemblance et on peut montrer que $\mathbb{E}[h^2(s, f_{\hat{\theta}_{\mathrm{mle}}})] > 0.38$ (voir la Section 2.3 de Birgé (2006)). Contrairement à $\hat{s}$, l'estimateur du maximum de vraisemblance $f_{\hat{\theta}_{\mathrm{mle}}}$ ne se rapproche pas de s lorsque le nombre d'observations n augmente.

1.2.2. Modèle de translation. Dans l'exemple précédent, nous pouvions nous comparer à l'estimateur du maximum de vraisemblance. Mais il existe des modèles très simples où cet estimateur n'existe pas. Prenons par exemple $\mathscr{F} = \{f_\theta,\, \theta \in [-1,1]\}$ où

$$f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}}\mathbb{1}_{[-1,1]}(x-\theta) & \text{pour tout } x \in \mathbb{R}\setminus\{\theta\} \\ 0 & \text{pour } x = \theta. \end{cases} \tag{1}$$

La méthode du maximum de vraisemblance ne permet pas de déterminer un estimateur consistant pour ce modèle. En revanche, utiliser une procédure basée sur des tests permet de construire l'estimateur $\hat{s}$ ci-dessous.

Proposition 2. *Il existe un estimateur $\hat{s} = f_{\hat{\theta}}$ de s tel que*

$$C\mathbb{E}\left[h^2(s, f_{\hat{\theta}})\right] \le \inf_{\theta\in[-1,1]} h^2(s, f_\theta) + n^{-1}$$

où f_θ est définie par (1) et où $C > 0$ est une constante universelle.

Comme pour le modèle de la loi uniforme, cet estimateur est robuste. On peut en outre montrer lorsque le modèle est vrai, c'est-à-dire lorsqu'il existe θ_0 tel que $s = f_{\theta_0}$, que l'estimateur $\hat{\theta}$ est consistent et converge vers θ_0 à la vitesse n^{-2}. Nous renvoyons à l'article de Birgé (2006) ou au Chapitre 4 pour une preuve de ce résultat.

1.3. Deux exemples en estimation non paramétrique. Ces procédures peuvent également être utilisées pour construire des estimateurs pour des modèles non paramétriques.

1.3.1. Hypothèses de régularité. En estimation non paramétrique, on estime habituellement s sous certaines hypothèses de régularité. Nous allons donc considérer un espace de fonctions régulières, et regarder les bornes de risque que l'on peut obtenir lorsque s appartient à cet espace.

Nous introduisons pour tout $p \in (0,+\infty]$ et tout $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d) \in (0,+\infty)^d$ l'espace

$$\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d)) = \begin{cases} \mathscr{B}^{\boldsymbol{\sigma}}_{\infty}(\mathbb{L}^p([0,1]^d)) & \text{si } p \in (0,1] \\ \mathscr{B}^{\boldsymbol{\sigma}}_{p}(\mathbb{L}^p([0,1]^d)) & \text{si } p \in (1,2) \\ \mathscr{B}^{\boldsymbol{\sigma}}_{\infty}(\mathbb{L}^p([0,1]^d)) & \text{si } p \in [2,+\infty) \\ \mathcal{H}^{\boldsymbol{\sigma}}([0,1]^d) & \text{si } p = \infty. \end{cases}$$

Dans la définition ci-dessus, $\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d))$ désigne l'espace de Besov (éventuellement anisotrope inhomogène) des fonctions définies sur $[0,1]^d$ de régularité $\boldsymbol{\sigma}$. Nous renvoyons à Hochmuth (2002) ou Akakpo (2009) pour une définition précise des espaces de Besov. L'espace $\mathcal{H}^{\boldsymbol{\sigma}}([0,1]^d)$ est celui des fonctions $\boldsymbol{\sigma}$-hölderiennes sur $[0,1]^d$.

Nous notons par $|\cdot|_{p,\boldsymbol{\sigma}}$ la semi-norme associée à l'espace $\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d))$ et par $\bar{\boldsymbol{\sigma}}$ la moyenne harmonique de $\boldsymbol{\sigma}$

$$\frac{1}{\bar{\boldsymbol{\sigma}}} = \frac{1}{d}\sum_{i=1}^{d} \frac{1}{\sigma_i}.$$

Nous pouvons obtenir une borne de risque lorsque $\sqrt{s}$ appartient à l'espace

$$\mathscr{B}([0,1]^d) = \bigcup_{p\in(0,+\infty]} \left(\bigcup_{\substack{\boldsymbol{\sigma}\in(0,+\infty)^d \\ \bar{\boldsymbol{\sigma}} > d(1/p-1/2)_+}} \mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d)) \right).$$

Plus précisément :

Théorème 3. *Il existe un estimateur $\hat{s}$ tel que si $\sqrt{s} \in \mathscr{B}([0,1]^d)$ alors*

$$C\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] \leq \left|\sqrt{s}\right|_{p,\boldsymbol{\sigma}}^{2d/(2\bar{\boldsymbol{\sigma}}+d)} n^{-2\bar{\boldsymbol{\sigma}}/(2\bar{\boldsymbol{\sigma}}+d)} + n^{-1}$$

où $p \in (0,+\infty]$, $\boldsymbol{\sigma} \in (0,+\infty)^d$, $\bar{\boldsymbol{\sigma}} > d(1/p-1/2)_+$ sont tels que $\sqrt{s} \in \mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d))$ et où $C > 0$ ne dépend que de d,p,$\boldsymbol{\sigma}$.

L'estimateur ainsi construit ne dépend que des observations X_i et non pas de la régularité $\boldsymbol{\sigma}$ de $\sqrt{s}$, qui est supposée inconnue. L'estimateur est donc adaptatif par rapport à $\boldsymbol{\sigma}$. Il est intéressant de noter que ce résultat est vrai pour tout $p > 0$ et tout $\boldsymbol{\sigma} \in (0,+\infty)^d$ tel que $\bar{\boldsymbol{\sigma}} > d(1/p-1/2)_+$, ce qui est, à notre connaissance, l'apanage des estimateurs construits par tests. Ce théorème est la version multidimensionnelle du Théorème 7 de Birgé (2006).

Remarque. L'hypothèse de régularité porte sur $\sqrt{s}$ et non sur s, ce qui s'explique par le fait que nous utilisons une perte Hellinger, qui correspond à une distance $\mathbb{L}_2$ entre les racines carrées des densités.

1.3.2. Hypothèses structurelles. La vitesse précédente est intéressante lorsque d est petit, mais devient lente lorsque d est grand. Ce phénomène est connu comme étant "le fléau de la

dimension" et il faut considérer d'autres types d'hypothèses sur s pour obtenir des vitesses plus rapides.

Il est possible de construire un estimateur par tests $\hat{s}$ sous des hypothèses très variées sur s. Un exemple simple donné à titre illustratif est celui où s est de la forme $s(x) = g(< \theta, x >)$ où g est une fonction régulière inconnue, θ un vecteur inconnu de $\mathbb{R}^d$ et $< \cdot, \cdot >$ le produit scalaire usuel de $\mathbb{R}^d$. On peut prouver (Corollaire 2 de Baraud et Birgé (2011)) :

Théorème 4. *Soit $\mathscr{F}$ l'ensemble des fonctions positives f définies sur $[-1,1]^d$ pour lesquelles il existe une fonction hölderienne g sur $[-1,1]$ et un vecteur θ de la boule unité ℓ_1 de $\mathbb{R}^d$ tels que $\sqrt{f(x)} = g(< \theta, x >)$ pour tout $x \in [-1,1]^d$, c'est-à-dire*

$$\mathscr{F} = \bigcup_{\sigma>0} \left\{ f,\, \exists g \in \mathcal{H}^\sigma([-1,1]),\, \exists \theta \in \mathbb{R}^d,\, \sum_{i=1}^d |\theta_i| \leq 1,\, \forall x \in [-1,1]^d,\, \sqrt{f(x)} = g(< \theta, x >) \right\}.$$

Il existe un estimateur $\hat{s}$ tel que si $s \in \mathscr{F}$, alors

$$C\mathbb{E}\left[h^2(s,\hat{s})\right] \leq |g|^{2/(2\sigma+1)}_{\infty,\sigma}\, n^{-2\sigma/(2\sigma+1)} + \frac{d \max\left\{1, \log n, \log\left(d^{-1}\,|g|^2_{\infty,\sigma\wedge 1}\right)\right\}}{n}.$$

Dans cette inégalité, g et σ sont tels que $g \in \mathcal{H}^\sigma([-1,1])$ et tels qu'il existe θ dans la boule unité ℓ^1 de $\mathbb{R}^d$ tel que pour tout $x \in [-1,1]^d$, $\sqrt{s(x)} = g(< \theta, x >)$. La constante $C > 0$ ne dépend que de σ.

Ce théorème montre que lorsque $\sqrt{s(x)} = g(< \theta, x >)$, la vitesse de convergence de $\hat{s}$ vers s est $n^{-2\sigma/(2\sigma+1)}$, ce qui correspond à la vitesse d'estimation de g. Elle est en particulier indépendante de la dimension d. En outre, l'estimateur est adaptatif par rapport à la régularité de g car cette vitesse est atteinte sans la connaissance *a priori* de σ. On peut également remarquer que cette vitesse est valable sans restriction sur σ.

1.4. Mélanger les hypothèses : un exemple. Il n'y a pas de dichotomie entre modèle paramétrique et modèle non paramétrique, et il est possible de travailler simultanément avec les deux à la fois.

Par exemple, lorsque s appartient au modèle paramétrique

$$\mathscr{F} = \left\{\theta^{-1}\mathbb{1}_{[0,\theta]},\, \theta > 0\right\},$$

la Proposition 1 assure que l'on peut construire un estimateur qui converge vers s à la vitesse n^{-1}. En revanche, cet estimateur est très mauvais lorsque s est loin du modèle paramétrique. Dans ce cas, si la fonction est régulière (au sens où $\sqrt{s} \in \mathscr{B}^\sigma([0,1])$ pour un $\sigma > 0$ inconnu), il est plus intéressant d'utiliser l'estimateur fournit par le Théorème 3 car il converge vers s à la vitesse $n^{-2\sigma/(2\sigma+1)}$.

Aucun des deux estimateurs n'est systématiquement meilleur que l'autre. Si s est suffisamment proche du modèle $\mathscr{F}$, l'estimateur donné par la Proposition 1 a un risque inférieur à celui donné par le Théorème 3. En revanche, si s est loin de $\mathscr{F}$, mais est régulière, l'estimateur donné par le Théorème 3 est meilleur.

Une solution est alors d'utiliser la remarque en dessous du Théorème 7 de Birgé (2006) pour construire l'estimateur $\hat{s}$ de la proposition suivante.

Proposition 5. *Il existe un estimateur $\hat{s}$ tel que*

$$C\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{\theta>0}\left\{h^2(s,\theta^{-1}\mathbb{1}_{[0,\theta]}) + \frac{\max\left\{\log\left(|\log\theta|/\Gamma_n\right),1\right\}}{n}\right\}$$

où $C>0$ est une constante universelle et où Γ_n est donné dans la Proposition 1. De plus, si $\sqrt{s}\in\mathscr{B}([0,1])$ alors

$$C'\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] \leq \left|\sqrt{s}\right|_{p,\sigma}^{2/(2\sigma+1)} n^{-2\sigma/(2\sigma+1)} + n^{-1}$$

où $p\in(0,+\infty]$, $\sigma\in((1/p-1/2)_+,+\infty)$, sont tels que $\sqrt{s}\in\mathscr{B}^\sigma(\mathbb{L}^p([0,1]))$ et où $C'>0$ ne dépend que de p,σ.

Lorsque s appartient au modèle paramétrique $\mathscr{F}$, le risque de l'estimateur satisfait $\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] = \mathcal{O}(n^{-1})$. Si s est éloigné de $\mathscr{F}$, mais est régulière, nous retrouvons la vitesse non paramétrique $\mathbb{E}\left[h^2\left(s,\hat{s}\right)\right] = \mathcal{O}(n^{-2\sigma/(2\sigma+1)})$.

Remarque. Cette propriété d'adaptation a un léger coût : les constantes C et C' qui apparaissent dans ce résultat sont plus petites que celles de la Proposition 6 et du Théorème 3.

2. Un résultat de sélection de modèles

Le but de cette section est d'expliquer plus précisément les résultats que l'on peut démontrer avec ces estimateurs.

2.1. La dimension métrique.

Ce qui permet de travailler avec des modèles paramétriques et non paramétriques est la notion de dimension métrique dans le sens de la Définition 6 de Birgé (2006).

Définition 1. *Soit $(\mathcal{L},d)$ un espace métrique, V un sous-ensemble de $\mathcal{L}$, et D_V un réel plus grand que $1/2$. Nous disons que V a une dimension métrique finie D_V si pour tout $\eta>0$, il existe $S_V(\eta)\subset\mathcal{L}$ tel que pour tout $f\in V$, il existe $g\in S_V(\eta)$ avec $d(f,g)\leq\eta$ et tel que*

$$\forall\varphi\in\mathcal{L},\ \forall x\geq 2,\quad |S_V(\eta)\cap\mathcal{B}(\varphi,x\eta)| \leq \exp\left(D_V\,x^2\right). \tag{2}$$

Dans l'inégalité ci-dessus, $\mathcal{B}(\varphi,x\eta)$ est la boule fermée centrée en φ de rayon $x\eta$ et $|S_V(\eta)\cap\mathcal{B}(\varphi,x\eta)|$ représente le cardinal de l'ensemble $S_V(\eta)\cap\mathcal{B}(\varphi,x\eta)$.

Un ensemble V admet donc une dimension métrique finie s'il possède de "bonnes" propriétés de discrétisation. L'ensemble $S_V(\eta)$ peut s'interpréter comme étant une version discrétisée de V. Cette version est proche de V puisque tout point $f\in V$ est à distance de $S_V(\eta)$ inférieure à η. La condition (2) demande que $S_V(\eta)$ ne contienne pas trop de points dans des boules. Le paramètre D_V règle justement ce nombre de points et s'interprète donc comme étant une mesure de la "masse" de l'ensemble V.

De nombreux exemples de modèles de dimension métrique finie peuvent être trouvés dans l'article de Birgé (2006). Donnons en deux.

Modèles paramétriques. Lorsque V est un ensemble de densités paramétré par un segment Θ de $\mathbb{R}$, on peut montrer le résultat suivant (Proposition 10 de Birgé (2006)).

Proposition 6. *Soit Θ un segment de $\mathbb{R}$ et V un sous-ensemble de $\mathbb{L}^1_{dens}(\mathbb{R}^d)$ de la forme*

$$V = \{f_\theta,\, \theta \in \Theta\}\,.$$

Supposons qu'il existe trois constantes c_1, c_2, α strictement positives telles que

$$c_1 \left|\theta - \theta'\right|^\alpha \leq h^2\left(f_\theta, f_{\theta'}\right) \leq c_2 \left|\theta - \theta'\right|^\alpha \quad \text{pour tout } \theta, \theta' \in \Theta.$$

Alors, V a une dimension métrique finie dans l'espace métrique $(\mathbb{L}^1_{dens}(\mathbb{R}^d), h)$ et

$$D_V = \max\left\{\alpha^{-1}\left(\log(c_2/c_1) + 2\log 5\right), 1/2\right\}.$$

De nombreux exemples de modèles V vérifiant l'hypothèse de cette proposition peuvent être trouvés dans le Chapitre 4. Par exemple, cette hypothèse est vraie avec $\alpha = 1$ pour le modèle uniforme $\left\{\theta^{-1}\mathbb{1}_{[0,\theta]},\, \Theta\right\}$ où Θ est un segment de $(0, +\infty)$. Elle est vraie avec $\alpha = 1/2$ pour le modèle

$$\mathscr{F} = \{f_\theta,\, \theta \in [-1,1]\} \quad \text{où} \quad f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}}\mathbb{1}_{[-1,1]}(x-\theta) & \text{pour tout } x \in \mathbb{R}\setminus\{\theta\} \\ 0 & \text{pour } x = \theta. \end{cases}$$

En outre, lorsque le modèle est suffisamment régulier, l'hypothèse est vraie avec $\alpha = 2$ (voir le Théorème 7.6 du Chapitre 1 de Ibragimov et Has'minskii (1981)).

Espaces vectoriels. Un autre exemple est celui où V est un sous-espace vectoriel de dimension finie de l'espace $\mathbb{L}^2(\mathbb{R}^d)$ des fonctions de carré intégrable sur $\mathbb{R}^d$ par rapport à la mesure de Lebesgue. Dans ce cas, si l'espace est non réduit à un singleton, la dimension métrique de V coïncide avec la dimension vectorielle.

Proposition 7. *Soit V un sous-espace vectoriel de dimension finie de $\mathbb{L}^2(\mathbb{R}^d)$. Alors, V a une dimension métrique finie dans $(\mathbb{L}^2(\mathbb{R}^d), d_2)$ bornée par $D_V = \max\{\dim V, 1/2\}$. Ici, d_2 est la distance standard de l'espace $\mathbb{L}^2(\mathbb{R}^d)$.*

Cette proposition découle de la preuve de la Proposition 8 de Birgé (2006).

2.2. Estimation sur un modèle. Pour tout ensemble V de dimension métrique finie, l'article de Birgé (2006) montre qu'il est possible d'utiliser une procédure basée sur des tests pour construire l'estimateur $\hat{s}$ du théorème suivant.

Théorème 8. *Soit V un ensemble de dimension métrique finie D_V dans l'espace métrique $(\mathbb{L}^1_{dens}(\mathbb{R}^d), h)$ ou dans $(\mathbb{L}^2(\mathbb{R}^d), d_2)$. Si V est de dimension métrique finie dans $\mathbb{L}^1_{dens}(\mathbb{R}^d)$, on note $B_V = \inf_{f\in V} h^2(s,f)$ tandis que s'il l'est dans $\mathbb{L}^2(\mathbb{R}^d)$, on note $B_V = \inf_{f\in V} d_2^2(\sqrt{s}, f)$. Il existe un estimateur $\hat{s}$ tel que pour tout $\xi > 0$,*

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq B_V + \frac{D_V}{n} + \xi\right] \leq e^{-n\xi},$$

où C est une constante strictement positive. En particulier,

$$C'\mathbb{E}\left[h^2(s,\hat{s})\right] \leq B_V + \frac{D_V}{n},$$

où C' est une constante strictement positive.

Si $V = \{f_\theta,\, \theta \in \Theta\}$ est un modèle paramétrique satisfaisant l'hypothèse de la Proposition 6, on déduit de ce résultat que l'estimateur $\hat{s}$ est tel que

$$C'\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{\theta\in\Theta} h^2(s,f_\theta) + \frac{\max\left\{\alpha^{-1}\left(\log(c_2/c_1) + 2\log 5\right), 1/2\right\}}{n}.$$

En particulier, si le modèle est correct, c'est-à-dire, si $s \in V$, alors $\mathbb{E}\left[h^2(s,\hat{s})\right] = \mathcal{O}(n^{-1})$. En revanche, si $\inf_{\theta\in\Theta} h^2(s,f_\theta) \leq n^{-1}$, ce qui signifie que le modèle est légèrement incorrect, le risque de l'estimateur $\hat{s}$ est toujours tel que $\mathbb{E}\left[h^2(s,\hat{s})\right] = \mathcal{O}(n^{-1})$. Cela s'interprète comme étant une propriété de robustesse.

Puisque ce théorème est valide pour les espaces de dimension métrique finie, on peut également l'appliquer à des sous-espaces vectoriels V de dimension finie de $\mathbb{L}^2(\mathbb{R}^d)$. On a alors

$$C'\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{f\in V} d_2^2(\sqrt{s},f) + \frac{\dim V}{n}.$$

Le premier terme du côté droit de cette inégalité est le terme de biais tandis que le second est celui du terme de variance. Idéalement, il faudrait trouver un modèle V pour lequel les deux termes sont petits afin de minimiser le risque de l'estimateur. Cela n'est pas évident car le terme de biais dépend des propriétés de s qui sont généralement inconnus. En outre, choisir V grand permet certes de diminuer le terme de biais, mais augmente malheureusement le terme de variance. Réciproquement, choisir V petit diminue le terme de variance, mais augmente le terme de biais. L'art de la sélection de modèles consiste, à partir des données et d'une liste $\mathbb{V}$ de modèles, d'en déterminer un qui réalise un bon compromis biais-variance.

2.3. Sélection de modèles. L'idée de la sélection de modèles remonte aux années 1970 avec les travaux pionniers de Akaike (1973) et Mallows (1973) pour des procédures basées sur la minimisation d'un contraste pénalisé. Des bornes de risque non-asymptotiques pour des estimateurs de ce type peuvent être trouvés dans les articles plus récents de Birgé et Massart (1997); Barron *et al.* (1999); Massart (2003). L'avantage des estimateurs construits par tests est qu'ils permettent, à l'heure actuelle, d'obtenir des théorèmes de sélection de modèles plus généraux que ceux obtenus par minimisation d'un contraste pénalisé.

Théorème 9. *Soit $\mathbb{V}$ une collection au plus dénombrable de modèles V de dimension métrique finie D_V dans l'espace $(\mathbb{L}^1_{dens}(\mathbb{R}^d), h)$ ou $(\mathbb{L}^2(\mathbb{R}^d), d_2)$. Si V est de dimension métrique finie dans $\mathbb{L}^1_{dens}(\mathbb{R}^d)$, on note $B_V = \inf_{f\in V} h^2(s,f)$ tandis que s'il l'est dans $\mathbb{L}^2(\mathbb{R}^d)$, on note $B_V = \inf_{f\in V} d_2^2(\sqrt{s},f)$. Soit alors Δ une application positive sur $\mathbb{V}$ telle que*

$$\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1.$$

Il existe un estimateur $\hat{s}$ tel que pour tout $\xi > 0$,

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq \inf_{V\in\mathbb{V}}\left\{B_V + \frac{D_V + \Delta(V)}{n}\right\} + \xi\right] \leq e^{-n\xi},$$

où C est une constante strictement positive. En particulier,

$$C'\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}}\left\{B_V + \frac{D_V + \Delta(V)}{n}\right\},$$

où C' est une constante strictement positive.

Ce résultat est particulièrement flexible et est valide sous des hypothèses faibles sur la collection de modèles $\mathbb{V}$. Cette collection peut inclure des modèles paramétriques ainsi que n'importe quel espace vectoriel de dimension finie. En outre, il ne fait pas intervenir la divergence de Kullback contrairement au Théorème 2 de Barron *et al.* (1999) pour les estimateurs du maximum de vraisemblance pénalisé. Comme pour le théorème précédent, ce résultat est issu de l'article de Birgé (2006).

Un exemple d'application est celui où $\mathbb{V}$ est une collection d'espaces vectoriels de dimension finie non nulle de $\mathbb{L}^2(\mathbb{R}^d)$. Dans ce cas,

$$C'\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}}\left\{\inf_{f\in V} d_2^2(\sqrt{s},f) + \frac{\dim V + \Delta(V)}{n}\right\}.$$

L'application Δ peut s'interpréter comme étant une (sous) probabilité sur la collection de modèles $\mathbb{V}$, permettant de travailler avec des familles très grandes de modèles (éventuellement infinies). Si l'on peut choisir $\Delta(V)$ de l'ordre de $\dim V$, ce qui signifie que la famille $\mathbb{V}$ ne contient pas trop de modèles par dimension, l'estimateur $\hat{s}$ réalise le meilleur compromis possible (à une constante multiplicative près) entre le terme de biais et celui de variance :

$$C''\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}}\left\{\inf_{f\in V} d_2^2(\sqrt{s},f) + \frac{\dim V}{n}\right\}. \tag{3}$$

Un autre exemple est celui où l'on souhaite travailler avec un modèle paramétrique $V = \{f_\theta,\, \theta\in\Theta\}$ indexé par un intervalle Θ non borné. Un tel ensemble n'est généralement pas de dimension métrique finie (il ne vérifie pas les hypothèses de la Proposition 6). Une solution est alors de décomposer V en une union d'ensembles de dimension métrique finie. Plus précisément, on écrit $\Theta = \cup_{i=1}^{\infty}\Theta_i$ comme une union de segments et on définit $V_i = \{f_\theta,\, \theta\in\Theta_i\}$. Lorsque les V_i sont de dimension métrique finie, on peut appliquer le théorème avec $\mathbb{V} = \{V_i,\, i\in\mathbb{N}^\star\}$ et par exemple $\Delta(V_i) = 2\log(i+1)$. On obtient alors un estimateur $\hat{s}$ tel que

$$C'''\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{\theta\in\Theta}\left\{h^2(s,f_\theta) + \frac{D_{V_{i(\theta)}} + \log\left(i(\theta)+1\right)}{n}\right\}$$

où $i(\theta)$ est n'importe quel entier de $\mathbb{N}^\star$ tel que $\theta\in V_{i(\theta)}$.

2.4. De la sélection de modèles à l'estimation. Le théorème précédent crée un lien entre statistiques et théorie de l'approximation dont le rôle est de trouver des modèles V de dimension "petite" possédant de "bonnes" propriétés d'approximation. Ces espaces sont parfois appelés des sieves (voir Grenander (1981); Birgé et Massart (1998)).

De nombreuses collections d'espaces vectoriels peuvent être utilisées pour déduire de (3) des bornes de risque sous des hypothèses convenables sur $\sqrt{s}$. Si l'on est intéressé par des hypothèses de régularité sur $\sqrt{s}$, la collection qui permet d'obtenir le Théorème 3 est celle proposée par Akakpo (2012). Les espaces vectoriels V de cette collection sont des espaces vectoriels de fonctions polynomiales par morceaux sur des partitions irrégulières anisotropes issues d'un algorithme d'approximation. Ces partitions sont la version anisotrope des partitions de DeVore et Yu (1990). C'est l'absence d'hypothèse sur les espaces vectoriels (autre que la dimension finie) qui permet d'obtenir les vitesses de convergence sur les espaces $\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^d))$ sous la seule condition que $p > 0$ et $\bar{\boldsymbol{\sigma}} > d(1/p - 1/2)_+$.

Un article taillé sur mesure pour fournir des collections $\mathbb{V}$ permettant de majorer le côté droit de (3) sous des hypothèses très variées sur $\sqrt{s}$ est celui de Baraud et Birgé (2011). Le Théorème 4 découle simplement de (3) appliqué avec une collection $\mathbb{V}$ proposée par Baraud et Birgé (2011). Le lecteur pourra trouver dans cet article des collections $\mathbb{V}$ permettant d'obtenir des vitesses de convergence lorsque $\sqrt{s}$ est la composée de fonctions régulières, lorsque $\sqrt{s}$ appartient à un "multiple index model" et bien d'autres.

3. Construction des estimateurs

Dans les sections précédentes, nous avons présenté des théorèmes qui assuraient l'existence d'estimateurs possédant de bonnes propriétés statistiques. Le but de cette section est d'expliquer comment on peut les construire.

Il existe plusieurs moyens de les construire et le choix dans cette introduction fut de considérer le test particulier de Baraud (2011) et d'utiliser la procédure générique de Birgé (2006). En accord avec la terminologie introduite par Lucien Birgé, ces estimateurs sont alors appelés des T-estimateurs (T pour test).

3.1. Sélectionner parmi deux fonctions. Pour comparer deux densités distinctes f et f', nous allons construire un test, c'est-à-dire une fonction mesurable des observations $\psi(\{f, f'\})$ renvoyant la fonction que l'on préfère.

Pour le définir, considérons la fonctionnelle introduite par Baraud (2011),

$$T(f,f') = \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{f'(X_i)} - \sqrt{f(X_i)}}{\sqrt{f(X_i) + f'(X_i)}} + \frac{1}{2}\int_{\mathbb{R}^d} \sqrt{f(x) + f'(x)}\left(\sqrt{f'(x)} - \sqrt{f(x)}\right) \mathrm{d}x$$

où la convention $0/0 = 0$ est utilisée (lorsque $f(X_i) = f'(X_i) = 0$). Cette fonctionnelle provient d'une formule variationnelle pour l'affinité de Hellinger et nous renvoyons à la Section 2 de Baraud (2011) pour les heuristiques qui conduisent à cette définition. Elles suggèrent que l'on

devrait préférer f' à f lorsque $T(f,f')>0$ et f à f' lorsque $T(f,f')<0$. Posons donc

$$\psi(\{f,f'\}) = \begin{cases} f' & \text{si } T(f,f')>0 \\ f & \text{si } T(f,f')<0 \end{cases}$$

et définissons $\psi(\{f,f'\})$ arbitrairement lorsque $T(f,f')=0$.

Cela donne une procédure statistique permettant de sélectionner parmi deux fonctions. Notons $\hat{f}\in\{f,f'\}$ la fonction que le test préfère. Alors, nous pouvons déduire de l'article de Baraud (2011) la proposition ci-dessous.

Proposition 10. *Pour tout $\xi>0$,*

$$\mathbb{P}\left[Ch^2(s,\hat{f}) \geq \inf\left\{h^2(s,f),h^2(s,f')\right\}+\xi\right] \leq e^{-n\xi}$$

où $C>0$ est une constante universelle.

Remarque. Un test naturel est celui du rapport de vraisemblance, et on peut se demander pourquoi nous ne l'utilisons pas. La réponse est qu'il ne possède pas les propriétés statistiques permettant d'obtenir la Proposition 10 sans hypothèse sur s, f et f'.

Afin d'estimer sur un modèle (comme le fait le Théorème 8), il faut étendre ce résultat. Cela se fait progressivement. Nous allons tout d'abord présenter une procédure permettant de sélectionner parmi une famille finie de points.

3.2. Sélectionner parmi une famille de points. Soit S un ensemble fini de densités. Nous allons utiliser le test pour comparer deux à deux les points de S.

Une situation très favorable est celle pour laquelle il existe une fonction $f\in S$ telle que $\psi(\{f,f'\})=f$ pour tout $f'\in S$, c'est-à-dire une fonction f que le test préfère à toutes les autres fonctions f' de S. Dans ce cas, on peut sélectionner f. Malheureusement, cette situation n'a aucune raison de se produire en général et il faut donc trouver un autre moyen de sélectionner parmi les fonctions de S.

Pour cela, introduisons pour toute fonction $f\in S$, l'ensemble $\mathcal{R}(f)$ des fonctions f' que le test préfère à f. Plus précisément :

$$\mathcal{R}(f) = \left\{f'\in S,\ f'\neq f,\ \psi(\{f,f'\})=f'\right\}.$$

Notons alors par $\mathcal{D}(f)$ le carré de la distance de Hellinger maximale entre f et les fonctions que le test lui préfère

$$\mathcal{D}(f) = \begin{cases} \sup\left\{h^2(f,f'),\ f'\in\mathcal{R}(f)\right\} & \text{si } \mathcal{R}(f)\neq\emptyset \\ 0 & \text{si } \mathcal{R}(f)=\emptyset. \end{cases}$$

Si $\mathcal{D}(f)$ est "grand", alors il existe une fonction f', préférée à f qui est loin de f. Cette fonction risque donc d'être un meilleur candidat à l'estimation de s que ne l'est f. En revanche, si $\mathcal{D}(f)$ est "petit", alors toute fonction f' préférée à f est proche de f. Tous les candidats meilleurs que f sont alors proches de f, et par conséquent f ne devrait pas être trop mauvais.

On va donc définir l'estimateur $\hat{s}$ comme étant n'importe quel estimateur à valeurs dans S tel que

$$\mathcal{D}(\hat{s}) = \inf_{f \in S} \mathcal{D}(f). \tag{4}$$

Il existe au moins un estimateur vérifiant (4) puisque S est finie. Il n'y a cependant pas unicité et plusieurs estimateurs peuvent vérifier (4).

Afin de comprendre les propriétés de cet estimateur, il faut remarquer que le test $\psi(\{f, f'\})$ entre f et f' peut faire une erreur : il peut préférer f' alors que $h^2(s, f)$ est bien plus petit que $h^2(s, f')$. Cela n'arrive qu'avec une probabilité faible, mais non nulle. Si l'on compare un grand nombre de fonctions, la probabilité que le test commette des erreurs n'est plus nécessairement négligeable et par conséquent, la "masse" de S interviendra dans la borne de risque. Ce qui permet de mesurer cette "masse" est la notion de D-modèles (D pour discret) au sens de la Définition 4 de Birgé (2006).

Définition 2. *Soit η et D deux réels strictement positifs. Un sous-ensemble S de $\mathbb{L}^1_{dens}(\mathbb{R}^d)$ est appelé D-modèle de paramètres η, D et 1 si*

$$|S \cap \mathcal{B}(f, x\eta)| \leq \exp\left[Dx^2\right] \quad \textit{pour tout } x \geq 2 \textit{ et } f \in \mathbb{L}^1_{dens}(\mathbb{R}^d).$$

Dans cette inégalité, $\mathcal{B}(f, x\eta)$ est la boule fermée centrée en f de rayon $x\eta$ dans l'espace métrique $(\mathbb{L}^1_{dens}(\mathbb{R}^d), h)$ et $|S \cap \mathcal{B}(f, x\eta)|$ représente le cardinal de $S \cap \mathcal{B}(f, x\eta)$.

Remarque. On peut vérifier que tout ensemble fini S est un D-modèle de paramètres η, $(\log|S|)/4$ et 1. Cependant, il est souvent possible de faire des calculs plus fins. Par exemple, considérons $T = \left\{k\eta^2,\, k \in \mathbb{Z},\, |k| \leq (2\eta^2)^{-1}\right\}$ et le modèle de translation

$$S = \left\{\mathbb{1}_{[\theta-1/2, \theta+1/2]},\, \theta \in T\right\}.$$

Alors S est un D-modèle de paramètres η, 1 et 1. En effet, on peut remarquer que pour toute densité f, $S \cap \mathcal{B}(f, x\eta)$ peut être vide auquel cas $|S \cap \mathcal{B}(f, x\eta)| = 0$. Dans le cas contraire, il existe $\theta \in T$ tel que

$$|S \cap \mathcal{B}(f, x\eta)| \leq |S \cap \mathcal{B}(\mathbb{1}_{[\theta-1/2, \theta+1/2]}, 2x\eta)|.$$

En remarquant que pour tout $\theta, \theta' \in T$,

$$h^2\left(\mathbb{1}_{[\theta-1/2, \theta+1/2]}, \mathbb{1}_{[\theta'-1/2, \theta'+1/2]}\right) = |\theta' - \theta|$$

on déduit

$$|S \cap \mathcal{B}(f, x\eta)| \leq \left|\left\{\theta' \in T,\ |\theta' - \theta| \leq 4x^2\eta^2\right\}\right|.$$

On peut alors majorer cette inégalité par

$$|S \cap \mathcal{B}(f, x\eta)| \leq 8x^2 + 1 \leq e^{x^2} \quad \text{pour tout } x \geq 2,$$

ce qui montre que S est bien un D-modèle de paramètres η, 1 et 1.

Lorsque l'ensemble S est un D-modèle, l'estimateur $\hat{s}$ construit par la procédure précédente vérifie l'inégalité suivante (Théorème 3 de Birgé (2006) ou Théorème 2 de Baraud (2011)).

Proposition 11. *Il existe une constante universelle $a > 0$ telle que pour tout D-modèle S de paramètres η, $D \geq 1/2$ et 1 tels que $an\eta^2 \geq D$, n'importe quel estimateur $\hat{s}$ satisfaisant (4) vérifie pour tout $\xi > 0$,*

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq \inf_{f\in S} h^2(s,f) + \eta^2 + \xi\right] \leq e^{-n\xi}$$

où $C > 0$ est une constante universelle.

3.3. Construction d'un estimateur sur un modèle. Ce résultat permet de construire un estimateur sur un ensemble discret S. Si l'on souhaite construire un estimateur sur un ensemble non discret $V \subset \mathbb{L}^1_{\text{dens}}(\mathbb{R}^d)$, comme par exemple un modèle paramétrique, on procède en deux étapes. La première étape consiste à discrétiser le modèle V, ce qui conduit à un D-modèle S. La seconde étape consiste à appliquer la procédure de la section précédente à S.

Autrement dit, on pourra obtenir des résultats pour des modèles V possédant de "bonnes" propriétés de discrétisation. Plus précisément, on travaillera avec des modèles V de dimension métrique finie dont nous avons déjà donné la définition, mais nous la rappelons ci-dessous.

Définition 3. *Soit V un sous-ensemble de $\mathbb{L}^1_{dens}(\mathbb{R}^d)$, et D_V un réel plus grand que $1/2$. Nous disons que V a une dimension métrique finie D_V si pour tout $\eta > 0$, il existe $S_V(\eta) \subset \mathbb{L}^1_{dens}(\mathbb{R}^d)$ tel que pour tout $f \in V$, il existe $g \in S_V(\eta)$ avec $h(f,g) \leq \eta$ et tel que*

$$\forall \varphi \in \mathbb{L}^1_{dens}(\mathbb{R}^d),\ \forall x \geq 2, \quad |S_V(\eta) \cap \mathcal{B}(\varphi, x\eta)| \leq \exp\left(D_V\, x^2\right).$$

Un ensemble V possède une dimension métrique finie D_V si pour tout $\eta > 0$, il existe un réseau $S_V(\eta)$ de V de pas η qui est un D-modèle $S_V(\eta)$ de paramètres η, D_V et 1. Le fait que $S_V(\eta)$ soit un réseau assure que nous ne perdons pas trop lors de la discrétisation. Le fait que $S_V(\eta)$ soit un D-modèle permet d'appliquer la procédure de la section précédente.

Précisément, on peut procéder ainsi pour construire un estimateur sur V. On définit η par $an\eta^2 = D_V$ où a est la constante de la Proposition 11. On applique alors la procédure précédente au D-modèle $S_V(\eta)$. L'estimateur $\hat{s}$ ainsi construit vérifie

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq \inf_{f\in S_V(\eta)} h^2(s,f) + \frac{D_V}{an} + \xi\right] \leq e^{-n\xi}.$$

Or, puisque $S_V(\eta)$ est un réseau de V

$$\begin{aligned}\inf_{f\in S_V(\eta)} h^2(s,f) &\leq \left(\inf_{f\in V} h(s,f) + \eta\right)^2 \\ &\leq 2\left(\inf_{f\in V} h^2(s,f) + \frac{D_V}{an}\right).\end{aligned}$$

Par conséquent, on obtient

$$\mathbb{P}\left[C'h^2(s,\hat{s}) \geq \inf_{f\in V} h^2(s,f) + \frac{D_V}{n} + \xi\right] \leq e^{-n\xi}$$

où C' est la constante universelle $C' = C/\max(2, 3/a)$. Cela donne l'estimateur $\hat{s}$ qui apparait dans le Théorème 8 pour des ensembles de dimension métrique finie dans $\mathbb{L}^1_{\text{dens}}(\mathbb{R}^d)$.

Remarque 1. Cette procédure est entièrement explicite pourvu que la discrétisation du modèle V le soit. Par exemple, on peut construire explicitement l'estimateur pour le modèle

$$V = \left\{ \mathbb{1}_{[\theta-1/2,\theta+1/2]},\, \theta \in [-1/2, 1/2] \right\},$$

car le calcul de la section précédente montre que la Définition 3 est vérifiée avec

$$S_V(\eta) = \left\{ \mathbb{1}_{[k\eta^2-1/2,k\eta^2+1/2]},\, k \in \mathbb{Z},\, |k| \leq (2\eta^2)^{-1} \right\}.$$

De manière plus générale, il est possible de discrétiser de manière explicite les modèles paramétriques satisfaisant les hypothèses de la Proposition 6.

Remarque 2. La construction de l'estimateur pour des ensembles de dimension métrique finie dans $\mathbb{L}^2(\mathbb{R}^d)$ est un peu plus complexe. Elle requiert un argument permettant de "transformer" un D-modèle dans $(\mathbb{L}^2(\mathbb{R}^d), d_2)$ en un D-modèle dans $(\mathbb{L}^1_{\text{dens}}(\mathbb{R}^d), h)$. Malheureusement, dans la plupart des cas, ce nouveau D-modèle est de nature purement abstraite et impossible à construire en pratique. La procédure est alors dans ce cas seulement théorique. Nous renvoyons à la démonstration du Théorème 6 de Birgé (2006) pour plus de détails.

3.4. Sélection de modèles. Si l'on souhaite obtenir un résultat de sélection de modèles, il faut modifier la procédure afin de prendre en compte la complexité de la famille de modèles. Si la procédure précédente pouvait être vue comme étant une alternative à celles basées sur la minimisation d'un contraste sur un modèle, celle qui suit peut être vue comme étant une alternative à celles basées sur la minimisation d'un contraste pénalisé.

3.4.1. La procédure. Prenons une collection $\mathbb{S} = \{S_V, V \in \mathbb{V}\}$ de D-modèles S_V de paramètres η_V, $D_V \geq 1/2$ et 1 indexé par un ensemble $\mathbb{V}$ au plus dénombrable. Nous allons comparer les fonctions de $S = \cup_{V\in\mathbb{V}} S_V$ deux à deux.

Pour cela, nous considérons une application positive pen sur S et modifions le test $\psi(\{f, f'\})$. Nous le définissons désormais par

$$\psi(\{f, f'\}) = \begin{cases} f' & \text{si } T(f, f') > \text{pen}(f') - \text{pen}(f) \\ f & \text{si } T(f, f') < \text{pen}(f') - \text{pen}(f) \end{cases}$$

et arbitrairement en cas d'égalité.

Nous introduisons ensuite pour tout $f \in S$, l'ensemble $\mathcal{R}(f)$ des fonctions f' que ce test préfère à f

$$\mathcal{R}(f) = \left\{ f' \in S,\ f' \neq f,\ \psi(\{f, f'\}) = f' \right\}$$

ainsi que la distance maximale au carré $\mathcal{D}(f)$ entre f et les fonctions que le test lui préfère :

$$\mathcal{D}(f) = \begin{cases} \sup\left\{ h^2(f, f'),\ f' \in \mathcal{R}(f) \right\} & \text{si } \mathcal{R}(f) \neq \emptyset \\ 0 & \text{si } \mathcal{R}(f) = \emptyset. \end{cases}$$

Soit alors $\hat{s}$ n'importe quel estimateur à valeurs dans S tel que

$$\max\left\{ \mathcal{D}(\hat{s}), \text{pen}(\hat{s}) \right\} = \inf_{f\in S} \left[\max\left\{ \mathcal{D}(f), \text{pen}(f) \right\} \right]. \tag{5}$$

En toute généralité, il n'existe pas nécessairement d'estimateur $\hat{s}$ vérifiant (5). Cependant, nous verrons qu'il en existe toujours au moins un pour les ensembles S et pénalités pen auquels nous nous intéresserons.

3.4.2. Sélection de D-modèles. Le résultat est le suivant (Théorème 5 de Birgé (2006)).

Théorème 12. *Soit $\mathbb{S} = \{S_V, V \in \mathbb{V}\}$ une collection de D-modèles S_V de paramètres η_V, $D_V \geq 1/2$ et 1 indexé par un ensemble $\mathbb{V}$ au plus dénombrable et $S = \cup_{V\in\mathbb{V}} S_V$.*

Il existe deux constantes universelles $a, b > 0$ telles que si $an\eta_V^2 \geq D_V/5$ pour tout $V \in \mathbb{V}$, si

$$\sum_{V\in\mathbb{V}} \exp\left(-an\eta_V^2\right) \leq 1,$$

et si

$$\text{pen}(f) = b \inf\left\{\eta_V^2,\, V \in \mathbb{V},\, S_V \ni f\right\} \quad \textit{pour tout } f \in S,$$

alors il existe un estimateur $\hat{s}$ tel que (5) soit vraie et n'importe lequel d'entre eux vérifie

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq \inf_{V\in\mathbb{V}}\left\{\inf_{f\in S_V} h^2(s,f) + \eta_V^2\right\} + \xi\right] \leq e^{-n\xi} \quad \textit{pour tout } \xi > 0,$$

où C est une constante universelle.

Ce résultat montre qu'il est possible d'obtenir un résultat de sélection de modèles pour des D-modèles. Il y a deux conditions sur ces modèles. La première, $an\eta_V^2 \geq D_V/5$, porte sur chaque D-modèle et est similaire à celle de la Proposition 11. La seconde, $\sum_{V\in\mathbb{V}} \exp\left(-an\eta_V^2\right) \leq 1$ porte sur la collection de modèles. Elle assure que la famille $\mathbb{V}$ n'est pas trop complexe afin de pouvoir borner les erreurs faites par le test.

3.4.3. Sélection de modèles de dimension métrique finie. Nous pouvons utiliser la procédure précédente pour faire de la sélection de modèles avec des modèles de dimension métrique finie. En effet, prenons une collection $\mathbb{V}$ au plus dénombrable d'ensembles V de dimension métrique finie $D_V \geq 1/2$ et une application positive Δ sur $\mathbb{V}$ telle que $\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1$. Nous pouvons discrétiser chaque ensemble V pour construire un D-modèle S_V et appliquer la procédure de la section précédente à $S = \cup_{V\in\mathbb{V}} S_V$.

Plus précisément, définissons pour tout $V \in \mathbb{V}$,

$$\eta_V = \max\left\{\sqrt{\frac{D_V}{5na}}, \sqrt{\frac{\Delta(V)}{na}}\right\}.$$

Soit alors $S_V = S_V(\eta_V)$ un D-modèle de paramètres η_V, D_V et 1 tel que

$$\inf_{f\in S_V} h(s,f) \leq \inf_{f\in V} h(s,f) + \eta_V. \tag{6}$$

La procédure précédente appliquée à $S = \cup_{V\in\mathbb{V}} S_V$ fournit un estimateur $\hat{s}$ tel que

$$\mathbb{P}\left[Ch^2(s,\hat{s}) \geq \inf_{V\in\mathbb{V}}\left\{\inf_{f\in S_V} h^2(s,f) + \eta_V^2\right\} + \xi\right] \leq e^{-n\xi} \quad \text{pour tout } \xi > 0.$$

Si l'on utilise (6) et que l'on remplace η_V par sa définition, on obtient donc bien l'inégalité souhaitée

$$\mathbb{P}\left[C'h^2(s,\hat{s}) \geq \inf_{V\in\mathbb{V}}\left\{\inf_{f\in V} h^2(s,f) + \frac{D_V + \Delta(V)}{n}\right\} + \xi\right] \leq e^{-n\xi} \quad \text{pour tout } \xi > 0,$$

où C' est une nouvelle constante universelle.

4. Estimation pratique

Les estimateurs précédents possèdent des propriétés théoriques intéressantes et la question naturelle qui survient lorsque l'on travaille avec eux est la suivante : peut-on les utiliser en pratique ?

En général, les procédures de la section précédente calculent beaucoup de tests, ce qui les rendent difficiles à implémenter. Parfois la situation est encore pire puisque la discrétisation des modèles n'est même pas explicite. C'est la raison pour laquelle ces estimateurs sont habituellement considérés dans la littérature comme des "benchmarks" permettant de montrer quels résultats il est théoriquement possible d'obtenir. Il y a cependant deux cas particuliers où il est possible d'utiliser des tests pour construire des estimateurs en pratique.

Le premier cas particulier est celui de l'estimation sur un modèle paramétrique et le Chapitre 4 est consacré à ce problème. Nous proposerons dans ce chapitre une nouvelle procédure basée sur le test de Baraud (2011) pour construire un estimateur en calculant moins de tests.

Le second cas particulier est celui de la sélection d'estimateurs. Il n'y a que deux articles qui traitent de ce sujet avec des tests dans la littérature : Baraud et Birgé (2009) et Baraud (2011). Le premier est consacré à la sélection d'estimateurs constants par morceaux tandis que le second peut sélectionner parmi des familles plus générales d'estimateurs. Si l'on accepte de couper l'échantillon en deux, d'utiliser la première partie pour construire les estimateurs et la seconde pour sélectionner parmi eux, alors cela est également traité dans Birgé (2006) (Section 9). Le point clé est que ces procédures sont basées sur la comparaison deux à deux des estimateurs. Leur complexité est donc de l'ordre du carré du nombre d'estimateurs. Elles sont donc facilement implémentables lorsque que la famille ne contient pas trop d'estimateurs et seulement théoriques dans le cas contraire. Une procédure alternative est proposée dans le Chapitre 3 pour sélectionner, en un temps raisonnable, parmi une famille particulière mais de cardinal très grand d'estimateurs constants par morceaux (il s'agira d'estimateurs constants par morceaux dont les partitions seront issues de l'algorithme d'approximation de DeVore et Yu (1990)).

5. À propos des autres cadres statistiques

Nous nous sommes limités précédemment au problème de l'estimation d'une densité par sélection de modèles avec une perte Hellinger. Cependant, il est possible d'obtenir des résultats de sélection de modèles pour d'autres pertes comme par exemple la perte $\mathbb{L}_1$ ou $\mathbb{L}_2$ (voir le Théorème 8 de Birgé (2006) pour un résultat en perte $\mathbb{L}_1$ et Birgé (2013) pour la perte $\mathbb{L}_2$). Il est également possible de travailler dans d'autres cadres statistiques. Dans la littérature, les cadres statistiques suivants ont été étudiés :

- La régression bornée à design déterministe ou aléatoire dans Birgé (2006, 2012); Baraud (2011).

- Le problème de l'estimation de l'intensité d'un processus de Poisson dans Birgé (2007); Baraud (2011); Baraud et Birgé (2009). Birgé (2007) peut estimer directement sa mesure moyenne.

- Le modèle du bruit blanc Gaussien (à variance connue) dans Birgé (2006).

- L'estimation de la densité conditionnelle de Y sachant X à partir d'un échantillon (X_i, Y_i) à l'aide d'une perte Hellinger déterministe dans Birgé (2012). Si l'on souhaite utiliser une perte Hellinger aléatoire, les procédures de Birgé (2006); Baraud (2011) peuvent également être utilisées.

- L'estimation de la densité de transition d'une chaîne de Markov homogène dans Birgé (2012).

- L'estimation des moyennes de variables aléatoires positives dans Baraud et Birgé (2009); Baraud (2011). Dans ce cadre statistique, nous observons n variables aléatoires $X_1, \dots, X_n$ indépendantes et positives. Chaque variable aléatoire X_i est supposée admettre une moyenne s_i et être "suffisamment" concentrée autour de sa moyenne. Le but est d'estimer le vecteur $(s_1, \dots, s_n)$.

- L'estimation de la densité d'un processus déterminental dans Baraud (2013).

- L'estimation de la fonction de hasard instantanée pour des données censurées et l'estimation de l'intensité de transition d'un processus de Markov dans Baraud et Birgé (2009) (en construisant un estimateur constant par morceaux sur une partition aléatoire).

6. Présentation des chapitres

Les apports de cette thèse sont à la fois théoriques et pratiques. Nous démontrons de nouveaux résultats théoriques dans deux cadres statistiques différents (Chapitres 2 et 3) et proposons de nouvelles procédures dans les Chapitres 3 et 4 orientées vers la pratique.

6.1. Chapitre 2. Dans le deuxième chapitre, nous nous intéressons au problème de l'estimation des intensités de plusieurs processus de Poisson indépendants $N_1, \dots, N_n$ indexés par un même ensemble $\mathbb{T}$. Nous supposons que ces processus sont liés de la manière suivante : il existe n covariables $x_1, \dots, x_n \in \mathbb{X}$, telles que N_i admet une intensité s_i par rapport à une mesure μ qui est de la forme $s_i(\cdot) = s(\cdot, x_i)$. Le but est alors d'estimer la fonction s à l'aide des observations $(N_i, x_i)_{1 \leq i \leq n}$.

Ce cadre statistique généralise le problème de l'estimation de l'intensité d'un processus de Poisson (cas où $n = 1$) et inclut le problème de la régression poissonienne (ce qui correspond au cas où les processus de Poisson sont en réalité des variables aléatoires poissoniennes).

Pour évaluer la performance des estimateurs, nous utilisons la distance de type Hellinger H définie pour toutes fonctions positives et intégrables f, f' par

$$H^2(f, f') = \frac{1}{2n} \sum_{i=1}^{n} \int_{\mathbb{T}} \left(\sqrt{f(t, x_i)} - \sqrt{f'(t, x_i)} \right)^2 \mathrm{d}\mu(t).$$

Nous établissons un résultat général de sélection de modèles dans le même esprit que le Théorème 9. Un modèle sera un ensemble de fonctions possédant de "bonnes" propriétés de discrétisation, et plus précisément un ensemble de dimension métrique borné :

Définition. *Notons M la mesure $M = n^{-1}\sum_{i=1}^n \mu\otimes\delta_{x_i}$ et $(\mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M), d_2)$ l'espace métrique des fonctions sur $\mathbb{T}\times\mathbb{X}$ de carré intégrable par rapport à la mesure M.*

Soit $V \subset \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ et D_V une application continue à droite de $(0,+\infty)$ vers $[1/2,+\infty)$ telle que $D_V(\eta) = o(\eta^2)$ lorsque $\eta \to +\infty$. On dit que V a une dimension métrique bornée par D_V si pour tout $\eta > 0$, il existe un sous ensemble $S_V(\eta) \subset \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ tel que pour tout $f \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$, il existe $g \in S_V(\eta)$ avec $d_2(f,g) \le \eta$ et tel que

$$\forall \varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M),\ \forall x \ge 2, \quad |S_V(\eta) \cap \mathcal{B}(\varphi, x\eta)| \le \exp\left(D_V(\eta)\, x^2\right).$$

Dans cette inégalité, $\mathcal{B}(\varphi, x\eta)$ représente la boule de $(\mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M), d_2)$ centrée en φ de rayon $x\eta$.

Lorsque D_V est une constante, on dit que V a une dimension métrique finie bornée par D_V.

Le résultat principal du chapitre est le suivant : prenons une collection $\mathbb{V}$ au plus dénombrable de modèles V de dimension métrique bornée $D_V(\cdot)$ et Δ une application positive sur $\mathbb{V}$ telle que $\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \le 1$. Alors, on montre dans ce chapitre que l'on peut construire un estimateur $\hat{s}$ tel que pour tout $\xi > 0$

$$\mathbb{P}\left[CH^2(s,\hat{s}) \ge \inf_{V\in\mathbb{V}}\left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\} + \xi\right] \le e^{-n\xi},$$

où $C > 0$ est une constante universelle et où

$$\eta_V = \inf\left\{\eta > 0,\ \frac{D_V(\eta)}{\eta^2} \le n\right\}.$$

En particulier, cela implique que

$$C'\mathbb{E}\left[H^2(s,\hat{s})\right] \le \inf_{V\in\mathbb{V}}\left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\},$$

où $C' > 0$ est une autre constante positive universelle.

Un choix adéquat des modèles V permet de travailler avec le cas où $\mathbb{T}\times\mathbb{X} = [0,1]^k$, μ la mesure de Lebesgue et s est régulière au sens où $\sqrt{s}$ est hölderienne de régularité $\boldsymbol{\alpha} \in (0,+\infty)^k$ sur $[0,1]^k$. L'estimateur $\hat{s}$ atteint alors la vitesse attendue $n^{-2\bar{\boldsymbol{\alpha}}/(2\bar{\boldsymbol{\alpha}}+k)}$ où $\bar{\boldsymbol{\alpha}}$ est la moyenne harmonique de $\boldsymbol{\alpha}$ et cela même lorsque la régularité $\boldsymbol{\alpha}$ est proche de 0. En outre, l'estimateur est adaptatif par rapport à $\boldsymbol{\alpha}$.

Nous étudions également le cas où s est (au moins approximativement) une fonction produit $s(t,x) = f(t)g(x)$. Nous déduisons du théorème général de sélection de modèles un second théorème de sélection de modèles où cette fois les modèles sont utilisés pour approcher les fonctions f et g. En particulier, lorsque $\mathbb{T} = [0,1]^{k_1}$, $\mathbb{X} = [0,1]^{k_2}$, μ la mesure de Lebesgue et $s(t,x) = f(t)g(x)$ avec f et g régulières (au sens où $\sqrt{f}$ est $\boldsymbol{\alpha}$-hölderienne et $\sqrt{g}$ est $\boldsymbol{\beta}$-hölderienne), bien choisir les modèles permet de construire un estimateur atteignant la vitesse

$\max(n^{-2\bar{\boldsymbol{\alpha}}/(2\bar{\boldsymbol{\alpha}}+k_1)}, n^{-2\bar{\boldsymbol{\beta}}/(2\bar{\boldsymbol{\beta}}+k_2)})$ qui est plus rapide que celle obtenue sous une hypothèse de régularité pure sur $\sqrt{s}$. De plus, l'estimateur est adaptatif par rapport aux régularités de f et g. En raison de la faiblesse des hypothèses sur les modèles du second théorème de sélection de modèles, on peut également considérer des hypothèses diverses sur f et g (par exemple, on peut supposer que f est régulière et g est une fonction composée).

Enfin, un dernier cas sur lequel nous avons travaillé est celui pour lequel il existe un modèle paramétrique $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ tel que l'intensité s_i de chaque processus de Poisson appartient à $\mathcal{F}$. Cela revient à supposer qu'il existe θ_i tel que $s_i(\cdot) = f_{\theta_i}(\cdot)$ et nous démontrons alors un théorème de sélection de modèles pour estimer la fonction $i \mapsto \theta_i$. Nous pouvons alors préciser les modèles pour obtenir des bornes de risque sous différentes hypothèses sur $i \mapsto \theta_i$. Typiquement, cette application peut être constante par morceaux ou être de la forme $\theta_i = \theta(x_i)$ où θ est une application sur $\mathbb{X}$. A titre d'illustration, cela permet de traiter le cas où $\mathbb{T} = (0, 1]$, $\mathbb{X} = [0, 1]^k$, μ la mesure de Lebesgue et où $s(t, x)$ est de la forme $\sqrt{s(t, x)} = a(x)t^{b(x)}$ où a est $\boldsymbol{\alpha}$-hölderienne et b est $\boldsymbol{\beta}$-holderienne telle que $b(x) > -1/2$ pour tout x. Dans ce cas, nous construisons un estimateur qui atteint la vitesse $\max\{(\log n/n)^{2\bar{\boldsymbol{\alpha}}/(2\bar{\boldsymbol{\alpha}}+k)}, (\log n/n)^{2\bar{\boldsymbol{\beta}}/(2\bar{\boldsymbol{\beta}}+k)}\}$. Cette vitesse correspond, à un terme logarithmique près, à la pire des vitesses d'estimation entre celle de a et celle de b.

Ce chapitre illustre l'intérêt de ces théorèmes généraux de sélection de modèles puisqu'il montre comment le choix adéquat de modèles permet d'obtenir des bornes de risque sous des hypothèses variées sur la fonction à estimer. En outre, les estimateurs sont adaptatifs et robustes par rapport à ces hypothèses.

6.2. Chapitre 3. Ce chapitre est consacré à l'estimation de la densité de transition d'une chaîne de Markov homogène. Nous observons $n + 1$ variables aléatoires $X_0, \ldots, X_n$ à valeurs dans un ensemble $\mathbb{X}$ et supposons que pour tout $i \in \{0, \ldots, n-1\}$ la loi conditionnelle de X_{i+1} sachant $X_i = x$ admet une densité $s(x, \cdot)$ par rapport à une mesure μ. Le but est alors d'estimer la densité de transition s sur un sous-ensemble $A \subset \mathbb{X}^2$.

Nous présentons deux nouvelles procédures dites "data-driven" qui permettront d'estimer s sous des hypothèses particulièrement faibles sur la chaîne de Markov. Le chapitre est divisé en deux parties, chaque partie étant dédiée à une procédure.

6.2.1. Première partie. La première procédure est basée sur la sélection d'estimateurs constants par morceaux. Dans cette première partie, $\mathbb{X} = \mathbb{R}^d$, $A = [0, 1]^{2d}$. Nous considérons une famille d'estimateurs constants par morceaux $\{\hat{s}_m, m \in \mathcal{M}\}$ où la collection de partitions $\mathcal{M}$ est définie par l'algorithme itératif de DeVore et Yu (1990). Nous proposons un nouveau test inspiré de Baraud (2011) afin de sélectionner parmi ces estimateurs. La procédure que nous utilisons est un mélange entre une procédure basée sur la minimisation d'un contraste pénalisé et une approche basée sur des tests. Cette procédure est un nouveau moyen d'utiliser un test pour sélectionner parmi une telle famille d'estimateurs. Elle peut s'interpréter comme étant une version implémentable des procédures de sélection d'estimateurs de Baraud et Birgé (2009); Baraud (2011).

En utilisant une perte aléatoire de type Hellinger, nous montrons que l'estimateur sélectionné $\hat{s}_{\hat{m}}$

vérifie une inégalité de type oracle. Plus précisément, nous montrons que

$$C\mathbb{E}\left[H^2\big(s\mathbb{1}_{[0,1]^{2d}}, \hat{s}_{\hat{m}}\big)\right] \leq \inf_{m\in\mathcal{M}}\left\{\mathbb{E}\left[H^2\big(s\mathbb{1}_{[0,1]^{2d}}, V_m\big)\right] + \frac{|m|\log n}{n}\right\}$$

où $C > 0$ est une constante universelle, où H est définie par

$$H^2(f,f') = \frac{1}{2n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\left(\sqrt{f(X_i,y)} - \sqrt{f'(X_i,y)}\right)^2 \mathrm{d}\mu(y)$$

et où V_m est le cone des fonctions positives constantes par morceaux sur la partition m :

$$V_m = \left\{\sum_{K\in m} a_K \mathbb{1}_K,\, \forall K\in m,\, a_K \in [0,+\infty)\right\}.$$

Nous soulignons que cette inégalité est valide sans hypothèse sur la chaîne de Markov autre que l'homogénéité. La construction de l'estimateur fait certes intervenir une pénalité qu'il faut calibrer, mais cette dernière ne dépend pas de certains paramètres sur la chaîne de Markov habituellement inconnus. Lorsque μ est la mesure de Lebesgue, nous pouvons en déduire des vitesses de convergence uniformes sur des boules d'espaces de Besov qui peuvent être inhomogènes et avoir un indice de régularité petit. L'estimateur est adaptatif par rapport à cet indice de régularité, et les vitesses correspondent, à un terme logarithmique près, à celles usuelles sur ces classes de fonctions.

Nous présentons ensuite des simulations numériques afin d'évaluer plus précisément la performance de cet estimateur et nous nous comparons en particulier à la procédure de Akakpo et Lacour (2011) qui est basée sur la minimisation d'un contraste $\mathbb{L}_2$ pénalisé inspiré des moindres carrés.

6.2.2. Deuxième partie. La première procédure permet de construire un estimateur constant par morceaux sur une partition aléatoire bien choisie. Cela permet certes d'estimer convenablement les densités de transition irrégulières (lorsque $\sqrt{s}|_{[0,1]^{2d}} \in \mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$ avec $\sigma < 1$) mais cela conduit inéluctablement à des vitesses de convergence sous-optimales lorsque la densité de transition est plus régulière (lorsque $\sigma > 1$).

Nous proposons alors une deuxième procédure afin d'obtenir de meilleures vitesses de convergence sous des hypothèses de régularité ou structurelles sur s. Cette nouvelle procédure est essentiellement théorique car elle est difficile à implémenter en pratique. Elle permet en revanche de démontrer un théorème de sélection de modèles. Ce théorème sera vraie sous l'hypothèse que la loi de X_i admet une densité par rapport à une mesure ν connue qui peut être majorée par une constante κ (inconnue) indépendante de i.

Plus précisément, considérons une collection $\mathbb{V}$ d'espaces vectoriels V de l'espace métrique $(\mathbb{L}^2(A, \nu\otimes\mu), d_2)$ des fonctions de carré intégrable sur A par rapport à la mesure produit $\nu\otimes\mu$. Prenons $(\Delta(V))_{V\in\mathbb{V}}$ une famille de nombres positifs telle que $\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1$. Nous montrons dans le Chapitre 3 que l'on peut construire un estimateur $\hat{s}$ tel que

$$C\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \inf_{V\in\mathbb{V}}\left\{d_2^2\left(\sqrt{s}|_A, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n}\right\}$$

où $C > 0$ dépend seulement de κ. En outre, l'estimateur peut être construit en utilisant seulement les observations et les modèles.

Bien choisir les modèles permet d'en déduire des vitesses de convergence pour une grande famille d'espaces de Besov (éventuellement inhomogènes et anisotropes) lorsque $\mathbb{X} = \mathbb{R}^d$, $A = [0,1]^{2d}$ et μ est la mesure de Lebesgue. L'estimateur est adaptatif par rapport à la régularité de la densité de transition. Un autre choix de modèles permet d'obtenir des vitesses de convergence plus rapides pour les chaînes de Markov auto-régressives

$$X_{i+1} = g_1(X_i) + g_2(X_i)\varepsilon_i$$

où g_1 et g_2 sont des fonctions régulières (de régularité supposée inconnue) et où les ε_i sont des variables aléatoires i.i.d non observées.

6.3. Chapitre 4. Ce dernier chapitre utilise le test de Baraud (2011) pour constuire un estimateur paramétrique d'une densité s à partir d'observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées.

Considérons un modèle paramétrique $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$ paramétré par un rectangle $\Theta = \prod_{j=1}^d [m_j, M_j]$ de $\mathbb{R}^d$ pour lequel il existe des nombres strictement positifs $\alpha_1, \dots, \alpha_d$, $\underline{R}_1, \dots, \underline{R}_d$, $\overline{R}_1, \dots, \overline{R}_d$ tels que pour tout $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$, $\boldsymbol{\theta'} = (\theta'_1, \dots, \theta'_d) \in \prod_{j=1}^d [m_j, M_j]$

$$\sup_{j\in\{1,\dots,d\}} \underline{R}_j |\theta_j - \theta'_j|^{\alpha_j} \leq h^2\left(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta'}}\right) \leq \sup_{j\in\{1,\dots,d\}} \overline{R}_j |\theta_j - \theta'_j|^{\alpha_j}$$

où $h(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta'}})$ est la distance de Hellinger entre les deux densités $f_{\boldsymbol{\theta}}$ et $f_{\boldsymbol{\theta'}}$. Les méthodes utilisant des tests sont généralement basées sur la discrétisation de $\mathscr{F}$ en un ensemble discret $\mathscr{F}_{\text{dis}}$ et la comparaison deux à deux des fonctions de $\mathscr{F}_{\text{dis}}$. Si cela fournit des résultats théoriques (voir le Théorème 8 par exemple), le nombre de tests qu'elles calculent est de l'ordre de $|\mathscr{F}_{\text{dis}}|^2$. Malheureusement, ce cardinal est souvent très grand, ce qui rend la construction de l'estimateur assez délicate en pratique.

Nous proposons une nouvelle procédure pour construire un estimateur sur un modèle paramétrique en effectuant moins de tests. La complexité de cette procédure est indépendante du cardinal de $\mathscr{F}_{\text{dis}}$ qui peut donc être très grand. Elle dépend plutôt de d, $\underline{R}_j$, $\overline{R}_j$, α_j, m_j et M_j ainsi que d'un paramètre mesurant la précision à laquelle on souhaite calculer l'estimateur. Sa complexité croît lentement lorsqu'on augmente la précision mais elle dépend fortement de d, $\underline{R}_j$, $\overline{R}_j$, α_j, m_j, M_j. Elle est destinée en pratique aux modèles de petite dimension d pour lesquels les $\overline{R}_j/\underline{R}_j$, $M_j - m_j$ et $1/\alpha_j$ ne sont pas trop grand.

D'un point de vue théorique, l'estimateur que nous proposons possède des propriétés statistiques similaires aux estimateurs précédents construits par tests. Nous obtenons une borne de risque non-asymptotique pour des modèles pour lesquels la méthode du maximum de vraisemblance peut ne pas fonctionner. Nous montrons que l'estimateur converge à la bonne vitesse lorsque le modèle est correct, et est robuste dans le cas contraire.

Nous présentons ensuite des simulations numériques afin d'évaluer plus précisément la performance de l'estimateur. Elles mettent en lumière le lien qu'il y a entre cet estimateur et celui du maximum de vraisemblance : lorsque le modèle paramétrique est suffisamment régulier, lorsque

la densité s est dans ce modèle, et lorsque la discrétisation $\mathscr{F}_{\text{dis}}$ est très fine, nous observons que l'estimateur est très proche du maximum de vraisemblance (avec grande probabilité). Il est cependant robuste, contrairement à l'estimateur du maximum de vraisemblance.

Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó.

Akakpo, N. (2009). *Estimation adaptative par sélection de partitions en rectangles dyadiques*. Thèse de doctorat, Université Paris Sud.

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21:1–28.

Akakpo, N. et Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401.

Baraud, Y. (2013). Estimation of the density of a determinantal process. *Confluentes Mathematici*, 5(1):3–21.

Baraud, Y. et Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.

Baraud, Y. et Birgé, L. (2011). Estimating composite functions by model selection. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. To appear.

Barron, A., Birgé, L. et Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413.

Birgé, L. (2006). Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. *In Asymptotics : particles, processes and inverse problems*, volume 55 de *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2012). Robust tests for model selection. *In From Probability to Statistics and Back : High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, volume 9, pages 47–64. IMS Collections.

Birgé, L. (2013). Model selection for density estimation with $\mathbb{L}_2$-loss. *Probability Theory and Related Fields*, pages 1–42.

Birgé, L. et Massart, P. (1997). From model selection to adaptive estimation. *In Festschrift for Lucien Le Cam*, pages 55–87. Springer.

Birgé, L. et Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 4(3):329–375.

DeVore, R. et Yu, X. (1990). Degree of adaptive approximation. *Mathematics of Computation*, 55:625–635.

Grenander, U. (1981). *Abstract Inference*. Probability and Statistics Series. John Wiley & Sons.

Hochmuth, R. (2002). Wavelet characterizations for anisotropic besov spaces. *Applied and Computational Harmonic Analysis*, 12:179–208.

Ibragimov, I. et Has'minskii, R. (1981). *Statistical estimation–asymptotic theory*. Applications of mathematics. Springer-Verlag.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15(4):661–675.

Massart, P. (2003). *Concentration inequalities and model selection*, volume 1896 de *Lecture Notes in Mathematics*. Springer Berlin/Heidelberg. École d'été de Probabilités de Saint-Flour.

CHAPITRE 2

Model selection for Poisson processes with covariates

Abstract

We observe n inhomogeneous Poisson processes with covariates and aim at estimating their intensities. We assume that the intensity of each Poisson process is of the form $s(\cdot, x)$ where x is the covariate and where s is an unknown function. We propose a model selection approach where the models are used to approximate the multivariate function s. We show that our estimator satisfies an oracle-type inequality under very weak assumptions both on the intensities and the models. By using an Hellinger-type loss, we establish non-asymptotic risk bounds and specify them under several kind of assumptions on the target function s such as being smooth or a product function. Besides, we show that our estimation procedure is robust with respect to these assumptions.

1. Introduction

We consider n independent Poisson point processes N_i for $i = 1, \ldots, n$ indexed by the measurable space $(\mathbb{T}, \mathscr{T})$. For each i, we assume that the intensity of N_i with respect to some reference measure μ on $(\mathbb{T}, \mathscr{T})$ is of the form $s_i(\cdot) = s(\cdot, x_i)$ where x_i is a deterministic element of some measurable set $(\mathbb{X}, \mathscr{X})$ and s is a non-negative function on $\mathbb{T} \times \mathbb{X}$ satisfying

$$\forall i \in \{1, \ldots, n\}, \quad \int_{\mathbb{T}} s(t, x_i) \, \mathrm{d}\mu(t) < +\infty.$$

Typically, this corresponds to the modelling of the times of failure of n repairable systems where the reliability of each of them depends on external factors measured by some covariates $x_1, \ldots, x_n$, in which case $\mathbb{T}$ corresponds to an interval of time, say $[0, 1]$, and $\mathbb{X}$ to some compact subset of $\mathbb{R}^k$, say $[0, 1]^k$. Our aim is to estimate s from the observations of the pairs $(N_i, x_i)_{1 \leq i \leq n}$.

Let $\mathbb{L}_+^1(\mathbb{T} \times \mathbb{X}, M)$ be the cone of integrable and non negative functions on $(\mathbb{T} \times \mathbb{X}, \mathscr{T} \otimes \mathscr{X})$ equipped with the product measure $M = \mu \otimes \nu_n$ where $\nu_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$. In order to evaluate the risks of our estimators, we endow $\mathbb{L}_+^1(\mathbb{T} \times \mathbb{X}, M)$ with the Hellinger-type distance H defined

for $u, v \in \mathbb{L}_+^1(\mathbb{T} \times \mathbb{X}, M)$ by

$$\begin{aligned} 2H^2(u,v) &= \int_{\mathbb{T}\times\mathbb{X}} \left(\sqrt{u(t,x)} - \sqrt{v(t,x)}\right)^2 \mathrm{d}M(t,x) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{T}} \left(\sqrt{u(t,x_i)} - \sqrt{v(t,x_i)}\right)^2 \mathrm{d}\mu(t). \end{aligned}$$

Let $(\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M), d_2)$ be the metric space of functions f on $\mathbb{T} \times \mathbb{X}$ such that f^2 belongs to $\mathbb{L}_+^1(\mathbb{T} \times \mathbb{X}, M)$. Given a suitable collection $\mathbb{V}$ of models (i.e subsets of $\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M)$ which are not necessarily linear spaces) and a non-negative application Δ on $\mathbb{V}$ satisfying

$$\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1,$$

we build an estimator $\hat{s}$ whose risk $\mathbb{E}\left[H^2(s,\hat{s})\right]$ satisfies

$$C\mathbb{E}\left[H^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n} \right\}, \tag{1}$$

where C is an universal positive constant, $d_2\left(\sqrt{s}, V\right)$ is the $\mathbb{L}^2$-distance between $\sqrt{s}$ and V and $n\eta_V^2$ is the metric dimension (in a suitable sense) of V. We shall use this inequality in order to derive risk bounds for our estimator under smoothness or structural assumptions on the target function s.

In the literature, much attention has been paid to the problem of estimating the intensity of a Poisson process without covariates. Concerning estimation by model selection, Reynaud-Bouret (2003) dealt with the $\mathbb{L}^2$-loss, and provided a model selection theorem for a family of linear spaces V. Baraud and Birgé (2009) used the Hellinger distance and considered the case where the sets V consist of piecewise constants functions on a partition of $\mathbb{T}$. More general models were considered by Birgé (2007) allowing for V any subset with finite metric dimensions (in a suitable sense).

Our statistical setting includes that of Poisson regression. Indeed, if one observes$n$ independent random variables $Y_1, \ldots, Y_n$, such that Y_i obeys to a Poisson law with parameter $f(x_i)$, one can estimate f by setting $\mathbb{T} = \{0\}$, $\mu = \delta_0$ the Dirac measure on $\mathbb{T}$ and $N_i(\{0\}) = Y_i$. In this case, $s(0,\cdot) = f(\cdot)$ and estimating s amounts to estimating f. This last issue has been studied in Antoniadis and Sapatinas (2001), Antoniadis et al. (2001), Baraud (2011) and Krishnamurthy et al. (2010) among other references. For the particular cases of Poisson regression and estimating the intensity of a single Poisson process, our results recover those of Baraud (2011).

If we except these cases, statistical procedures that can estimate s from n independent Poisson processes with covariates are rather scarce. The only risk bounds we are aware of are due to Comte et al. (2011) who considered the $\mathbb{L}^2$-loss and penalized projection estimators on linear spaces. Their approach requires that the intensity s be bounded from above by a quantity that needs to be either known or suitably estimated. Besides, they impose some restrictions on the family of linear spaces V in order that their estimator possesses minimax properties over classes of functions which are smooth enough.

Our approach is based on robust testing. We propose a test inspired from a variational formula in Baraud (2011) and then apply the general methodology for model selection developed in Birgé (2006). This yields a T-estimator $\hat{s}$ that possesses nice (adaptation and robustness) properties but suffers from the fact that its construction is numerically intractable. This estimator should be thus considered as a benchmark for what theoretical feasible. We obtain an oracle inequality of the form (1) under very mild assumptions both on the intensity s and the family of models $\mathbb{V}$. This allows to derive risk bounds over a large range of Hölderian spaces including irregular ones.

We shall also consider functions s defined on a subset $\mathbb{T} \times \mathbb{X}$ of a linear space with large dimension, say $\mathbb{T} \times \mathbb{X} = [0,1]^{1+k}$ with a large value of k. It is well known that in such a situation, the minimax approach based on smoothness assumptions may lead to very slow rates of convergence. This phenomenon is known as the curse of dimensionality. In this case, an alternative approach is to assume that s belongs to classes $\mathscr{F}$ of functions satisfying structural assumptions (such as the multiple index model, the generalized additive model, the multiplicative model ...) and for which faster rates of convergence can be achieved. Very recently, this approach was developed by Juditsky et al. (2009) (for the Gaussian white noise model) and by Baraud and Birgé (2011) (in more general settings). Unlike Juditsky et al. (2009) we shall not assume that s belongs to $\mathscr{F}$ but rather consider $\mathscr{F}$ as an approximating class for s.

In this work, our point of view is closer to that developed in Baraud and Birgé (2011). We shall use our new model selection theorem in conjunction with suitable families $\mathbb{V}$ of models in order to design an estimator $\hat{s}$ possessing good statistical properties with respect to many classes of functions of interest, including classes $\mathscr{F} = \mathscr{F}_{\times}$ of product functions $(t,x) \mapsto u(t)v(x)$. When $s(t,x)$ is of the form (or close to) $u(t)v(x)$, where u and v are assumed to be smooth we shall prove that our estimator is fully adaptive with respect to the regularities of both u and v. We shall also consider structural assumptions on the functions u and v as well as parametric ones when t and x lie in a large dimensional space. We shall study the situation where, the intensity of each Poisson process belongs to a parametric class of functions $\mathscr{F}_{\Theta}$ with $\Theta \subset \mathbb{R}^k$. This means that there exists some element $f_{\theta(x_i)} \in \mathscr{F}_{\Theta}$ such that $s(\cdot, x_i) = f_{\theta(x_i)}(\cdot)$, and our aim is then to estimate the mapping $x \mapsto \theta(x)$ by model selection.

This chapter is organized as follows. The general model selection theorem can be found in Section 2. In Section 3, we study the case where $\mathscr{F}$ is a class of smooth functions, and in Section 4 the case where $\mathscr{F}$ is a class of product functions. The problem of estimating s when the intensity of each Poisson process N_i belongs to the same parametric model is dealt in Section 5. Section 6 is devoted to the proofs.

Let us introduce some notations that will be used all along the chapter. We set $\mathbb{N}^{\star} = \mathbb{N} \setminus \{0\}$, $\mathbb{R}^{\star} = \mathbb{R} \setminus \{0\}$. The components of a vector $\boldsymbol{\theta} \in \mathbb{R}^k$ are denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. The numbers $x \wedge y$ and $x \vee y$ stand for $\min(x,y)$ and $\max(x,y)$ respectively. For $(E, \mathcal{E}, \nu)$ a measured space, we denote by $\mathbb{L}^2(E,\nu)$ the linear space of measurable functions f such that $\int_E |f|^2 \, \mathrm{d}\nu < \infty$. When $(E,\nu) = (\mathbb{T} \times \mathbb{X}, M)$, the corresponding $\mathbb{L}^2$-distance is denoted by d_2, and the norm by $\|\cdot\|_2$. Alternatively, this distance (respectively this norm) is denoted by $d_{\mathtt{t}}$ (respectively $\|\cdot\|_{\mathtt{t}}$) when $(E,\nu) = (\mathbb{T}, \mu)$, and by $d_{\mathtt{x}}$ (respectively $\|\cdot\|_{\mathtt{x}}$) when $(E,\nu) = (\mathbb{X}, \nu_n)$. The supremum norm of a bounded function f on a domain E is denoted by $\|f\|_{\infty} = \sup_{x \in E} |f(x)|$. For (E,d) a metric space, $x \in E$ and $A \subset E$, the distance between x and A is denoted by $d(x,A) = \inf_{a \in A} d(x,a)$. The closed ball centered at $x \in E$ with radius r is denoted by $\mathcal{B}(x,r)$. The cardinality of a finite set A is denoted by $|A|$. We use $\mathscr{F}$ as a generic notation for a family of functions of $\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M)$

of special interest. The notations C,C',C''... are for constants. The constants C,C',C''... may change from line to line.

2. A general model selection theorem

Throughout this chapter, a model V is a subset of $\mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ with bounded metric dimension, in the sense of Definition 6 of Birgé (2006). We recall this definition below.

Definition 1. *Let V be a subset of $\mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ and D_V a right-continuous map from $(0,+\infty)$ into $[1/2,+\infty)$ such that $D_V(\eta) = o(\eta^2)$ when $\eta \to +\infty$. We say that V has a metric dimension bounded by D_V if for all $\eta > 0$, there exists $S_V(\eta) \subset \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ such that for all $f \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$, there exists $g \in S_V(\eta)$ with $d_2(f,g) \le \eta$ and such that*

$$\forall \varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M),\ \forall x \ge 2, \quad |S_V(\eta) \cap \mathcal{B}(\varphi, x\eta)| \le \exp\left(D_V(\eta)\, x^2\right).$$

Moreover, if one can choose D_V as a constant, we say that V has a finite metric dimension bounded by D_V.

This notion is more general than the dimension for linear spaces since a linear space V with finite dimension (in the usual sense) has a finite metric dimension. Besides, if V is not reduced to $\{0\}$ one can choose $D_V = \dim V$, what we shall do along this chapter. The link with the classical definition of metric entropy may be found in Section 6.4.3 of Birgé (2006). Other models of interest with bounded metric dimension will appear later in the chapter.

Given a collection of such subsets, our approach is based on model selection. We propose a selection rule based on robust testing in the spirit of the papers Birgé (2006); Baraud (2011). The test and the selection rule which are mainly abstract are postponed to Section 6. The main result is the following.

Theorem 1. *Let $\mathbb{V}$ be an at most countable family of models V with bounded metric dimension $D_V(\cdot)$ and Δ be a mapping from $\mathbb{V}$ into $[0,+\infty)$ such that*

$$\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \le 1.$$

There exists an estimator $\hat{s} \in \mathbb{L}^1_+(\mathbb{T}\times\mathbb{X}, M)$ such that, for all $\xi > 0$,

$$\mathbb{P}\left[CH^2(s,\hat{s}) \ge \inf_{V\in\mathbb{V}}\left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\} + \xi\right] \le e^{-n\xi},$$

where C is an universal positive constant and where

$$\eta_V = \inf\left\{\eta > 0,\ \frac{D_V(\eta)}{\eta^2} \le n\right\}.$$

In particular, by integrating the above inequality,

$$C'\mathbb{E}\left[H^2(s,\hat{s})\right] \le \inf_{V\in\mathbb{V}}\left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\}, \tag{2}$$

where C' is an universal positive constant.

The condition $\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1$ can be interpreted as a (sub)probability on the collection $\mathbb{V}$. The more complex the family $\mathbb{V}$, the larger the weights $\Delta(V)$. When $\mathbb{V}$ consists of linear spaces V of finite dimensions D_V one can take $\eta_V^2 = D_V/n$ and hence (2) leads to

$$C'\mathbb{E}\left[H^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ d_2^2\left(\sqrt{s}, V\right) + \frac{D_V + \Delta(V)}{n} \right\}.$$

When one can choose $\Delta(V)$ of order D_V, which means that the family $\mathbb{V}$ of models does not contain too many models per dimension, the estimator $\hat{s}$ achieves the best trade-off (up to a constant) between the approximation and the variance terms.

In the remaining part of this chapter, we shall consider subsets $\mathscr{F} \subset \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ corresponding to various assumptions on $\sqrt{s}$ (smoothness, structural, parametric assumptions ...). For such an $\mathscr{F}$, we associate a collection $\mathbb{V}_{\mathscr{F}}$ and deduce from Theorem 1 a risk bound for the estimator $\hat{s}$ whenever $\sqrt{s}$ belongs or is close to $\mathscr{F}$. This bound takes the form

$$C''\mathbb{E}\left[H^2(s,\hat{s})\right] \leq \inf_{f\in\mathscr{F}} \left\{ d_2^2(\sqrt{s}, f) + \varepsilon_{\mathscr{F}}(f) \right\} \tag{3}$$

where

$$\varepsilon_{\mathscr{F}}(f) = \inf_{V\in\mathbb{V}_{\mathscr{F}}} \left\{ d_2^2(f,V) + \eta_V^2 + \frac{\Delta(V)}{n} \right\},$$

and we shall bound the term $\varepsilon_{\mathscr{F}}(f)$ from above. This upper bound will mainly depend on some properties of f, for example smoothness ones. In this case, this result says that if $\sqrt{s}$ is irregular but sufficiently close to a smooth function f, the bound we get essentially corresponds to the one we would get for f. This can be interpreted as a robustness property.

Sometimes, several assumptions on $\sqrt{s}$ are plausible, and one does not know what class $\mathscr{F}$ should be taken. A solution is to consider a collection $\mathfrak{F}$ of such classes $\mathscr{F}$ and to use the proposition below to get an estimator whose risk satisfies (up to a remaining term) relation (3) simultaneously for all classes $\mathscr{F} \in \mathfrak{F}$.

Proposition 2. *Let $\mathfrak{F}$ be an at most countable collection of subsets of $\mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ and $\bar{\Delta}$ be a mapping on $\mathfrak{F}$ into $[0,+\infty)$ such that $\sum_{\mathscr{F}\in\mathfrak{F}} e^{-\bar{\Delta}(\mathscr{F})} \leq 1$. For all $\mathscr{F}\in\mathfrak{F}$, let $\mathbb{V}_{\mathscr{F}}$ be a collection of models and $\Delta_{\mathscr{F}}$ be a mapping such that the assumptions of Theorem 1 hold.*

There exists an estimator $\hat{s}$ such that, for all $\mathscr{F}\in\mathfrak{F}$,

$$C\mathbb{E}\left[H^2(s,\hat{s})\right] \leq \inf_{f\in\mathscr{F}} \left\{ d_2^2(\sqrt{s}, f) + \varepsilon_{\mathscr{F}}(f) \right\} + \frac{\bar{\Delta}(\mathscr{F})}{n},$$

where

$$\varepsilon_{\mathscr{F}}(f) = \inf_{V\in\mathbb{V}_{\mathscr{F}}} \left\{ d_2^2(f,V) + \eta_V^2 + \frac{\Delta_{\mathscr{F}}(V)}{n} \right\},$$

and where C is an universal positive constant.

3. Smoothness assumptions

Let $\mathbf{I} = \prod_{j=1}^k I_j$ where the I_j are intervals of $\mathbb{R}$ and $\boldsymbol{\alpha} = \boldsymbol{\beta} + \mathbf{p} \in (0,+\infty)^k$ with $\mathbf{p}\in\mathbb{N}^k$ and $\boldsymbol{\beta} \in (0,1]^k$. A function f belongs to the Hölder class $\mathcal{H}^{\boldsymbol{\alpha}}(\mathbf{I})$, if there exists $L(f) \in [0,+\infty)$ such that

for all $(x_1, \ldots, x_k) \in \mathbf{I}$ and all $j \in \{1, \ldots, k\}$, the functions $f_j(x) = f(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_k)$ admit a derivative of order p_j satisfying

$$\left| f_j^{(p_j)}(x) - f_j^{(p_j)}(y) \right| \leq L(f)|x - y|^{\beta_j} \quad \forall x, y \in I_j.$$

The class $\mathcal{H}^{\boldsymbol{\alpha}}(\mathbf{I})$ is said to be isotropic when the α_j are all equal, and anisotropic otherwise, in which case $\bar{\boldsymbol{\alpha}}$ given by $\bar{\boldsymbol{\alpha}}^{-1} = k^{-1} \sum_{j=1}^{k} \alpha_j^{-1}$ corresponds to the average smoothness of a function f in $\mathcal{H}^{\boldsymbol{\alpha}}(\mathbf{I})$. We define the class of Hölderian functions on $\mathbf{I}$ by

$$\mathcal{H}(\mathbf{I}) = \bigcup_{\boldsymbol{\alpha} \in (0, +\infty)^k} \mathcal{H}^{\boldsymbol{\alpha}}(\mathbf{I}).$$

Assuming that $\sqrt{s}$ is Hölderian corresponds thus to the choice $\mathscr{F} = \mathcal{H}(\mathbb{T} \times \mathbb{X})$. Anisotropic classes of smoothness are of particular interest in our context since the function s depends on variables t and x that may play very different roles.

Families of linear spaces possessing good approximation properties with respect to the elements of $\mathscr{F}$ can be found in the literature. We refer to the results of Dahmen et al. (1980). We may use these linear spaces (models) to approximate the elements of $\mathscr{F}$, and deduce from Theorem 1 the following result.

Corollary 1. *Let us assume that $\mathbb{T} \times \mathbb{X} = [0,1]^k$ and that μ is the Lebesgue measure. There exists an estimator $\hat{s}$ such that for all $f \in \mathcal{H}\left([0,1]^k\right)$,*

$$C\mathbb{E}\left[H^2(s, \hat{s})\right] \leq d_2^2(\sqrt{s}, f) + L(f)^{\frac{2k}{2\bar{\alpha}+k}} n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+k}} + n^{-1} \tag{4}$$

where $\boldsymbol{\alpha} \in (0, +\infty)^k$ is such that $f \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^k)$ and where $C > 0$ depends only on k and $\max_{1 \leq j \leq k} \alpha_j$.

Remark that the risk bound given by inequality (4) holds without any restriction on $\boldsymbol{\alpha}$. Such a generality can be obtained since our model selection theorem is valid for any collection $\mathbb{V}$ of finite dimensional linear spaces. Some restrictions on the dimensionality of the linear spaces $V \in \mathbb{V}$ (as in Comte et al. (2011)) would prevent us to get this rate of convergence for the Hölder classes $\mathcal{H}^{\boldsymbol{\alpha}}\left([0,1]^k\right)$ when $\min_{1 \leq j \leq k} \alpha_j$ is too small.

The preceding risk bound is quite satisfactory if k is small but becomes worse when k increases. We shall therefore consider other types of classes in the next section in order to avoid this curse of dimensionality.

4. Families $\mathscr{F}$ of product functions

A common way of modelling the influence of the covariates on the number of failures of n systems is to assume that, for each $i \in \{1, \ldots, n\}$, the intensity of N_i, is of the form $s(t, x_i) = u(t)v(x_i)$ where u is an unknown density function on $\mathbb{T}$, and v some unknown function from $\mathbb{X}$ into $[0, +\infty)$. This means, that in average, the number of failures of system i, $\mathbb{E}[N_i(\mathbb{T})] = v(x_i)$, depends on x_i through v only, and conditionally to $N_i(\mathbb{T}) = k_i > 0$, the times of failure are distributed along $\mathbb{T}$ independently of x_i, but accordingly to the density u.

We shall therefore consider the class $\mathscr{F}$ defined by

$$\mathscr{F} = \left\{\kappa v_1 v_2,\, \kappa \geq 0,\, (v_1, v_2) \in \mathbb{L}^2(\mathbb{T}, \mu) \times \mathbb{L}^2(\mathbb{X}, \nu_n),\, \|v_1\|_{\mathtt{t}} = \|v_2\|_{\mathtt{x}} = 1\right\}, \tag{5}$$

which amounts to assuming that s is of the form (or close to) a product function $u(t)v(x)$ with $u = v_1^2$ and $v = \kappa^2 v_2^2$.

In this section, we introduce collections of models $\mathbb{V}_1$ and $\mathbb{V}_2$ in order to approximate the components v_1 and v_2 separately. Given $V_1 \in \mathbb{V}_1$ to approximate v_1 and $V_2 \in \mathbb{V}_2$ to approximate v_2, we approximate $v_1 v_2$ by the model $V_1 \otimes V_2$ defined by

$$V_1 \otimes V_2 = \left\{v_1' v_2',\, (v_1', v_2') \in V_1 \times V_2\right\}. \tag{6}$$

The metric dimension of $V_1 \otimes V_2$ is controlled as follows.

Lemma 1. *Let V_1 and V_2 be a finite dimensional linear space of $\mathbb{L}^2(\mathbb{T}, \mu)$ and $\mathbb{L}^2(\mathbb{X}, \nu_n)$ respectively. The set $V_1 \otimes V_2$ defined by (6) has a finite metric dimension bounded by*

$$D_{V_1 \otimes V_2} = 1.4\,(\dim V_1 + \dim V_2 + 1).$$

By using Theorem 1, we prove the following result.

Proposition 3. *Let $\mathbb{V}_1$ (respectively $\mathbb{V}_2$) be an at most countable collection of finite dimensional linear spaces of $\mathbb{L}^2(\mathbb{T}, \mu)$ (respectively $\mathbb{L}^2(\mathbb{X}, \nu_n)$). Let, for all $i \in \{1, 2\}$, Δ_i be a non-negative mapping on $\mathbb{V}_i$ such that*

$$\sum_{V_i \in \mathbb{V}_i} e^{-\Delta_i(V_i)} \leq 1.$$

There exists an estimator $\hat{s}$ such that, for all $\kappa v_1 v_2 \in \mathscr{F}$, where $\mathscr{F}$ is defined by (5),

$$\begin{aligned} C\mathbb{E}\left[H^2(s, \hat{s})\right] \;\leq\; & d_2^2(\sqrt{s}, \kappa v_1 v_2) + \inf_{V_1 \in \mathbb{V}_1} \left\{\kappa^2 d_{\mathtt{t}}^2(v_1, V_1) + \frac{\dim V_1 \vee 1 + \Delta_1(V_1)}{n}\right\} \\ & + \inf_{V_2 \in \mathbb{V}_2} \left\{\kappa^2 d_{\mathtt{x}}^2(v_2, V_2) + \frac{\dim V_2 \vee 1 + \Delta_2(V_2)}{n}\right\} \end{aligned}$$

where C is an universal positive contant. Furthermore, $\sqrt{\hat{s}}$ belongs to $\mathscr{F}$.

Apart for the term $d_2^2(\sqrt{s}, \kappa v_1 v_2)$ which corresponds to some robustness with respect to the assumption $\sqrt{s} \in \mathscr{F}$, the risk bound we get corresponds to the one we would get if we could apply a model selection theorem on the components v_1 and v_2 separately.

4.1. Smoothness assumptions on v_1 and v_2. We illustrate this proposition by setting $\mathbb{T} = [0,1]^{k_1}$, $\mathbb{X} = [0,1]^{k_2}$, μ the Lebesgue measure and

$$\mathscr{F} = \left\{\kappa v_1 v_2, \kappa \geq 0, v_1 \in \mathcal{H}([0,1]^{k_1}), \|v_1\|_{\mathtt{t}} = 1, v_2 \in \mathcal{H}([0,1]^{k_2}), \|v_2\|_{\mathtt{x}} = 1\right\}. \tag{7}$$

We apply Proposition 3 with families $\mathbb{V}_1$ and $\mathbb{V}_2$ of linear spaces possessing good approximation properties with respect to the functions of $\mathcal{H}([0,1]^{k_1})$ and $\mathcal{H}([0,1]^{k_2})$ respectively. This leads to the following corollary.

Corollary 2. *There exists an estimator $\hat{s}$ such that, for all $\kappa v_1 v_2 \in \mathscr{F}$, where $\mathscr{F}$ is defined by (7),*

$$C\mathbb{E}\left[H^2(s,\hat{s})\right] \leq d_2^2(\sqrt{s},\kappa v_1 v_2) + \kappa^{\frac{2k_1}{2\bar{\alpha}+k_1}} L\,(v_1)^{\frac{2k_1}{2\bar{\alpha}+k_1}}\, n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+k_1}} + \kappa^{\frac{2k_2}{2\bar{\beta}+k_2}} L\,(v_2)^{\frac{2k_2}{2\bar{\beta}+k_2}}\, n^{-\frac{2\bar{\beta}}{2\bar{\beta}+k_2}} + n^{-1}$$

where $\boldsymbol{\alpha} \in (0,+\infty)^{k_1}$, is such that $v_1 \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_1})$, where $\boldsymbol{\beta} \in (0,+\infty)^{k_2}$ is such that $v_2 \in \mathcal{H}^{\boldsymbol{\beta}}([0,1]^{k_2})$, and where $C>0$ depends only on k_1, k_2, $\max_{1\leq j\leq k_1}\alpha_i$, and $\max_{1\leq j\leq k_2}\beta_i$.

In particular, if s is a product function of the form $\sqrt{s} = \kappa v_1 v_2$ where $v_1 \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_1})$, and $v_2 \in \mathcal{H}^{\boldsymbol{\beta}}([0,1]^{k_2})$, $\sqrt{s}$ is Hölderian with regularity $(\boldsymbol{\alpha},\boldsymbol{\beta})$ on $[0,1]^{k_1+k_2}$. However, the rate given by the corollary above is always faster than the one we would get by Corollary 1 under smoothness assumption only.

4.2. Mixing smoothness and structural assumptions. When k_2 is large, we may consider structural assumptions on v_2 instead of smoothness ones to improve the risk bound. Proposition 3 allows to consider a wide variety of situations thanks to the approximation results of Baraud and Birgé (2011) on composite functions. We do not present all of them for the sake of concisely. We just consider the example in which the class $\mathscr{F}$ is

$$\begin{aligned}\mathscr{F} = \Big\{ & \kappa v_1 v_2,\ \kappa \geq 0,\ v_1 \in \mathcal{H}([0,1]^{k_1}),\ \boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_l \in \mathcal{B}(0,1), g \in \mathcal{H}([-1,1]^l), \\ & \forall \boldsymbol{x} \in \mathbb{X},\ v_2(\boldsymbol{x}) = g\left(<\boldsymbol{\theta}_1,\boldsymbol{x}>,\ldots,<\boldsymbol{\theta}_l,\boldsymbol{x}>\right),\ \|v_1\|_{\mathtt{t}} = \|v_2\|_{\mathtt{x}} = 1\Big\}\end{aligned} \tag{8}$$

where $\mathbb{T} = [0,1]^{k_1}$, μ is the Lebesgue measure and

$$\mathbb{X} = \mathcal{B}(0,1) = \left\{ \boldsymbol{x} \in \mathbb{R}^{k_2},\ \sum_{j=1}^{k_2} x_j^2 \leq 1 \right\}$$

is the unit ball of $\mathbb{R}^{k_2}$. The following corollary ensues from Proposition 3 and Corollary 2 of Baraud and Birgé (2011).

Corollary 3. *There exists an estimator $\hat{s}$ such that, for all $\kappa v_1 v_2 \in \mathscr{F}$, where $\mathscr{F}$ is defined by (8),*

$$C\mathbb{E}\left[H^2\left(s,\hat{s}\right)\right] \leq d_2^2(\sqrt{s},\kappa v_1 v_2) + \kappa^{\frac{2k_1}{2\bar{\alpha}+k_1}} L(v_1)^{\frac{2k_1}{2\bar{\alpha}+k_1}}\, n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+k_1}} + \kappa^{\frac{2l}{2\bar{\beta}+l}} L(g)^{\frac{2l}{2\bar{\beta}+l}}\, n^{-\frac{2\bar{\beta}}{2\bar{\beta}+l}} + \frac{\ln\left(\kappa^2\|g\|_{\boldsymbol{\beta}}^2 k_2^{-1}\right) \vee \ln n \vee 1}{n} k_2$$

where $\boldsymbol{\alpha} \in (0,+\infty)^{k_1}$, $\boldsymbol{\beta} \in (0,+\infty)^{l}$ are such that $v_1 \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_1})$, $g \in \mathcal{H}^{\boldsymbol{\beta}}([-1,1]^l)$ with $v_2(\boldsymbol{x}) = g\left(<\boldsymbol{\theta}_1,\boldsymbol{x}>,\ldots,<\boldsymbol{\theta}_l,\boldsymbol{x}>\right)$ and where $C>0$ depends only on k_1, l, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In the above inequality, $\|g\|_{\boldsymbol{\beta}}$ stands for any positive real number such that for all $(x_1,\ldots,x_l) \in [-1,1]^l$ and all $j \in \{1,\ldots,l\}$, the function $g_j(x) = g(x_1,\ldots,x_{j-1},x,x_{j+1},\ldots,x_l)$ satisfies

$$|g_j(x) - g_j(y)| \leq \|g\|_{\boldsymbol{\beta}} |x-y|^{\beta_j \wedge 1} \quad \forall x,y \in [-1,1].$$

When $\sqrt{s}$ belongs to the class $\mathscr{F}$, the risk bound of the above inequality corresponds to the one we would get if we could estimate the functions v_1 and g separately. This risk bound is then better than the one we would get under smoothness assumptions on v_2 when $l < k_2$.

4.3. Examples of parametric assumptions. Theorem 1 also allows to deal with parametric assumptions. Hereafter, we consider a class $\mathscr{F}$ of the form

$$\mathscr{F} = \left\{ a u_b v_{\boldsymbol{\theta}},\, a \geq 0,\, b \in I,\, \boldsymbol{\theta} \in \mathbb{R}^{k_2} \right\},$$

where I is an interval of $\mathbb{R}$, $(u_b)_{b\in I}$ is a family of functions and $v_{\boldsymbol{\theta}}$ is defined by $v_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left(< \boldsymbol{x}, \boldsymbol{\theta} >\right)$ for $\boldsymbol{x} \in \mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^{k_2}, \sum_{j=1}^{k_2} x_j^2 \leq 1\}$, the unit ball of $\mathbb{R}^{k_2}$. For each $i \in \{1, \dots n\}$, the intensity of N_i is thus assumed to be proportional to an element of (or an element close to) some reference parametric model $\{u_b^2,\, b \in I\}$. Let us give 3 examples of such models.

The Power Law Processes are Poisson processes whose intensities are proportional to $u_b(t) = t^b$ for all $t \in \mathbb{T} = (0, 1]$ and some $b \in (-1/2, +\infty)$. Proposed first in Duane (1964), this model is popular in reliability. Indeed, although the intensity is simple, different situations can be modelled by this model. For example, if $b = 0$ each N_i obeys to an homogeneous Poisson process, whereas if $b > 0$ (respectively $b < 0$) the reliability of each system reduces (respectively improves) with time. In software reliability, we can cite the Goel-Okumoto model of Goel and Okumoto (1979) and the S-Shaped model of Yamada et al. (1983). The former considers intensities proportional to $u_b(t) = e^{-bt}$ whereas the latter corresponds to $u_b(t) = \sqrt{t}e^{-bt}$ where $b \in [0, +\infty)$ and $t \in \mathbb{T} = [0, +\infty)$.

We consider the following assumption on the family $\{u_b,\, b \in I\}$.

Assumption 1. *The family $(u_b)_{b\in I}$ is a family of non vanishing functions of $\mathbb{L}^2(\mathbb{T}, \mu)$ indexed by an interval I of the form $(b_0, +\infty)$. Moreover, there exists two positive non-increasing functions $\underline{\rho}, \bar{\rho}$ on I, such that for all $b, b' \in I$,*

$$\underline{\rho}\left(b \vee b'\right) |b - b'| \leq \left\| \frac{u_b}{\|u_b\|_{\mathtt{t}}} - \frac{u_{b'}}{\|u_{b'}\|_{\mathtt{t}}} \right\|_{\mathtt{t}} \leq \bar{\rho}\left(b \wedge b'\right) |b - b'|.$$

The purpose of the lemmas below is to show that the above assumption holds for the Duane, Goel-Okumoto and S-Shaped models.

Lemma 2. *Let $I = (-1/2, +\infty)$, $\mathbb{T} = (0, 1]$, μ the Lebesgue measure, and for $b \in I$, $u_b(t) = t^b$. Assumption 1 is satisfied with*

$$\underline{\rho}(u) = \bar{\rho}(u) = \frac{1}{1 + 2u} \quad \textit{for all } u > -1/2.$$

Lemma 3. *Let $I = (0, +\infty)$, $\mathbb{T} = [0, +\infty)$, μ the Lebesgue measure, $k \in \mathbb{N}$, and for $b \in I$, $u_b(t) = t^{k/2}e^{-bt}$. Assumption 1 is satisfied with*

$$\underline{\rho}(u) = \frac{1}{2u} \quad \textit{and} \quad \bar{\rho}(u) = \frac{\sqrt{k+1}}{2u} \quad \textit{for all } u > 0.$$

All along this section, $\|\cdot\|$ denotes the standard Euclidean norm of $\mathbb{R}^{k_2}$

$$\forall \boldsymbol{x} \in \mathbb{R}^{k_2}, \quad \|\boldsymbol{x}\|^2 = \sum_{j=1}^{k_2} x_j^2$$

and d the distance induced by this norm.

Proposition 4. *Let $(u_b)_{b\in I}$ be a family such that Assumption 1 holds. There exist $\hat{a}\geq 0$, $\hat{b}\in I$ and $\hat{\boldsymbol{\theta}}\in\mathbb{R}^{k_2}$, such that the estimator $\hat{s}=(\hat{a}u_{\hat{b}}v_{\hat{\boldsymbol{\theta}}})^2$ satisfies, for all $a\geq 0$, $b\in I$, $\boldsymbol{\theta}\in\mathbb{R}^{k_2}$, and $f\in\mathscr{F}$ of the form $f(t,\boldsymbol{x})=au_b(t)v_{\boldsymbol{\theta}}(\boldsymbol{x})$,*

$$C\mathbb{E}\left[H^2(s,\hat{s})\right]\leq d_2^2\left(\sqrt{s},f\right)+\frac{k_2\left(1\vee\|\boldsymbol{\theta}\|\right)}{n}+\frac{C'}{n} \tag{9}$$

where C is an universal positive constant and where C' depends only on $\underline{\rho}$, $\bar{\rho}$, b_0 and b. More precisely,

$$C'=\log\left[1\vee\bar{\rho}\left(b_0+\frac{b-b_0}{b-b_0+1}\right)\right]+\left|\log\left(1\wedge\underline{\rho}(1+b)\right)\right|+\left|\log(b-b_0)\right|.$$

Under parametric assumptions on s, this result says that the rate of convergence of $\hat{s}$ is of order n^{-1}, which is quite satisfying when n is large, but may be inadequate in a non-asymptotic point of view. Indeed, the second term of the right-hand side of inequality (9) may be large especially when k_2 is large, says larger than n. This difficulty can be overcome by considering that $\boldsymbol{\theta}$ is sparse, which means that $\boldsymbol{\theta}$ is close to some (unknown) linear subspace W of $\mathbb{R}^{k_2}$ with $\dim W$ small. Below, we generalize Proposition 4 to take account of this situation.

Proposition 5. *Let $(u_b)_{b\in I}$ be a family such that Assumption 1 holds. Let $\mathbb{W}$ be an at most countable family of linear subspaces of $\mathbb{R}^{k_2}$ and let Δ be a non-negative map on $\mathbb{W}$ such that $\sum_{W\in\mathbb{W}}e^{-\Delta(W)}\leq 1$.*

There exist $\hat{a}\geq 0$, $\hat{b}\in I$ and $\hat{\boldsymbol{\theta}}\in\mathbb{R}^{k_2}$, such that the estimator $\hat{s}=(\hat{a}u_{\hat{b}}v_{\hat{\boldsymbol{\theta}}})^2$ satisfies, for all $a\geq 0$, $b\in I$, $\boldsymbol{\theta}\in\mathbb{R}^{k_2}$, and $f\in\mathscr{F}$ of the form $f(t,\boldsymbol{x})=au_b(t)v_{\boldsymbol{\theta}}(\boldsymbol{x})$,

$$\begin{aligned}C\mathbb{E}\left[H^2(s,\hat{s})\right] &\leq d_2^2(\sqrt{s},f)+\frac{C'}{n}\\ &\quad+\inf_{W\in\mathbb{W}}\left\{a^2\|u_b\|_{\mathfrak{t}}^2e^{2\|\theta\|}d^2(\boldsymbol{\theta},W)+\frac{(1\vee\dim W)(1\vee\|\boldsymbol{\theta}\|)+\Delta(W)}{n}\right\}\end{aligned}$$

where C is an universal positive constant and where C' is given by Proposition 4.

For illustration purpose, let us make explicit the constant C' for the Duane model, and let us therefore assume that there exist some unknown parameters $a,b,\boldsymbol{\theta}$ such that s is of the form $\sqrt{s(t,\boldsymbol{x})}=at^b\exp\left(<\boldsymbol{\theta},\boldsymbol{x}>\right)$. We derive from Proposition 4 an estimator whose risk satisfies

$$C\mathbb{E}\left[H^2(s,\hat{s})\right]\leq\frac{\left(1\vee\|\boldsymbol{\theta}\|\right)k_2+\left|\log(2b+1)\right|}{n} \tag{10}$$

where C is an universal positive constant. However, if for instance k_2 is large and if most of the components of $\boldsymbol{\theta}$ are small or null, the preceding proposition can be used to improve substantially the risk of our estimators. For simplicity, assume that

$$k_\star=\left|\{j\in\{1,\dots,k_2\},\theta_j\neq 0\}\right|$$

is small. We then define the set $\mathcal{M}$ of all subsets of $\{1,\dots,k_2\}$, and for each $m\in\mathcal{M}$, the set

$$W_m=\{(y_1,\dots,y_{k_2}),\,\forall j\notin m,\,y_j=0\}\subset\mathbb{R}^{k_2}.$$

We apply Proposition 5 with

$$\mathbb{W} = \{W_m,\, m \in \mathcal{M}\} \quad \text{and} \quad \forall m \in \mathcal{M},\ \Delta(W_m) = 1 + |m| + \log \binom{k_2}{|m|}.$$

This leads to an estimator $\hat{s}$ such that

$$C'' \mathbb{E}\left[H^2(s, \hat{s})\right] \leq \frac{(1 \vee \log k_2 \vee \|\boldsymbol{\theta}\|)(1 \vee k_\star) + |\log(2b+1)|}{n},$$

which improves inequality (10) when $k_\star$ is small and k_2 large.

5. Parametric models

In this section, we consider the natural situation where the intensity of each process N_i belongs (or is close) to a same parametric model. Throughout this section, $n \geq 2$. Let us consider a closed rectangle Θ of $\mathbb{R}^k$, that is a subset of $\mathbb{R}^k$ for which there exist $m_1, \ldots, m_k \in \mathbb{R} \cup \{-\infty\}$ and $M_1, \ldots, M_k \in \mathbb{R} \cup \{\infty\}$ such that

$$\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^k,\, \forall i \in \{1, \ldots, k\},\, m_i \leq x_i \leq M_i\right\}.$$

Let us denote by $\mathcal{F} = \{f_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$ a class of functions of $\mathbb{L}^2(\mathbb{T}, \mu)$. Our aim is to estimate s when, for each $i \in \{1, \ldots n\}$, the square root of the intensity of the Poisson process N_i, $\sqrt{s(\cdot, x_i)}$, is (or is close to) an element of $\mathcal{F}$. We introduce thus the class of functions $\mathscr{F}$ defined by

$$\mathscr{F} = \left\{(t, x) \mapsto f_{\boldsymbol{u}(x)}(t), \text{ where } \boldsymbol{u} \text{ is a map from } \mathbb{X} \text{ into } \Theta\right\}.$$

For instance, if $\mathcal{F}$ corresponds to the Duane model (see Section 4.3), Θ is a closed rectangle included in $\mathbb{R} \times (-1/2, +\infty)$ and

$$\mathcal{F} = \left\{at^b,\, (a, b) \in \Theta\right\}.$$

The class $\mathscr{F}$ is then the set of all functions f of the form $f(t, x) = a(x)t^{b(x)}$ where a and b are two functions on $\mathbb{X}$ such that $(a(x), b(x)) \in \Theta$ for all $x \in \mathbb{X}$.

We consider the following assumption to deal with more general classes $\mathcal{F}$.

Assumption 2. *The set Θ is a closed rectangle of $\mathbb{R}^k$. There exist $\boldsymbol{\alpha} = (\alpha_j)_{1 \leq j \leq k} \in (0, 1]^k$ and $\boldsymbol{R} = (R_j)_{1 \leq j \leq k} \in (0, +\infty)^k$ such that*

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta'} \in \Theta, \quad \|f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta'}}\|_{\mathfrak{t}} \leq \sum_{j=1}^{k} R_j \left|\theta_j - \theta'_j\right|^{\alpha_j}. \tag{11}$$

The aim of the lemmas below is to prove that this assumption is satisfied for the Duane, Goel-Okumoto and S-Shaped models.

Lemma 4. *Let μ be the Lebesgue measure, and for all $\boldsymbol{\theta} \in \mathbb{R} \times [-1/2, +\infty)$,*

$$f_{\boldsymbol{\theta}}(t) = \theta_1 t^{\theta_2} \quad \text{for all } t \in \mathbb{T} = (0, 1].$$

Then, for all positive numbers r_1, r_2, and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [-r_1, r_1] \times [-1/2 + 1/r_2, +\infty)$,

$$\|f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}'}\|_{\mathtt{t}} \leq r_2^{1/2} |\theta_1 - \theta_1'| + \sqrt{2} r_1 r_2^{3/2} |\theta_2 - \theta_2'|.$$

Lemma 5. *Let μ be the Lebesgue measure, and for all $k \in \{0, 1\}$, $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R} \times (0, +\infty)$,*

$$f_{\boldsymbol{\theta}}(t) = \theta_1 t^{k/2} e^{-\theta_2 t} \quad \text{for all } t \in \mathbb{T} = (0, +\infty).$$

Let r_1, r_2 be two positive numbers and let us set

$$\begin{array}{ll} C_1(0) = (r_2/2)^{1/2} & C_2(0) = r_1 r_2^{3/2}/2 \\ C_1(1) = r_2/2 & C_2(1) = (3/8)^{1/2} r_1 r_2^2. \end{array}$$

For all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [-r_1, r_1] \times [1/r_2, +\infty)$,

$$\|f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}'}\|_{\mathtt{t}} \leq C_1(k) |\theta_1 - \theta_1'| + C_2(k) |\theta_2 - \theta_2'|.$$

Remark. For the Duane model, Lemma 4 shows that Assumption 2 is fulfilled for $\Theta = [-r_1, r_1] \times [-1/2 + 1/r_2, +\infty)$. This will allow to obtain a risk bound when $\sqrt{s}$ is close to the class

$$\begin{aligned} \mathscr{F}_{r_1, r_2} \quad = \quad & \Big\{ (t, x) \mapsto a(x) t^{b(x)}, \text{where } a \text{ maps } \mathbb{X} \text{ into } [-r_1, r_1] \text{ and} \\ & \qquad b \text{ maps } \mathbb{X} \text{ into } [-1/2 + 1/r_2, +\infty) \Big\}. \end{aligned}$$

By using Proposition 2 with $\mathfrak{F} = \{\mathscr{F}_{r_1, r_2},\ r_1, r_2 \in \mathbb{N}^\star\}$, we can also derive a risk bound for the class

$$\begin{aligned} \mathscr{F} \quad &= \quad \bigcup_{r_1, r_2 \in \mathbb{N}^\star} \mathscr{F}_{r_1, r_2} \\ &= \quad \Big\{ (t, x) \mapsto a(x) t^{b(x)}, \text{where } a \text{ maps } \mathbb{X} \text{ into a compact subset of } \mathbb{R}, \\ & \qquad \text{and } b \text{ maps } \mathbb{X} \text{ into a closed interval included in } (-1/2, +\infty) \Big\}. \end{aligned}$$

5.1. A model selection theorem. The main theorem of Section 5 is the following.

Theorem 6. *Suppose that Assumption 2 holds. Let $\mathbb{W}_1, \ldots, \mathbb{W}_k$ be k families of finite dimensional linear subspaces of $\mathbb{L}^2(\mathbb{X}, \nu_n)$. Let for each $j \in \{1, \ldots, k\}$, Δ_j be a non-negative mapping on $\mathbb{W}_j$ such that $\sum_{W_j \in \mathbb{W}_j} e^{-\Delta_j(W_j)} \leq 1$.*

There exists an estimator $\hat{s}$ such that for all map $\boldsymbol{u} = (u_1, \ldots, u_k)$ from $\mathbb{X}$ with values into Θ, and $f \in \mathscr{F}$ of the form $f(t, x) = f_{\boldsymbol{u}(x)}(t)$,

$$C \mathbb{E}\left[H^2(s, \hat{s})\right] \leq d_2^2(\sqrt{s}, f) + \sum_{j=1}^{k} \varepsilon_j(u_j)$$

where $\varepsilon_j(u_j)$ is defined by

$$\varepsilon_j(u_j) = \inf_{W_j \in \mathbb{W}_j} \left\{ R_j^2 \ (d_{\mathbf{x}}(u_j, W_j))^{2\alpha_j} + \frac{(\dim(W_j) \vee 1)\, \tau_{\boldsymbol{u},j}(n) + \Delta_j(W_j)}{n} \right\},$$

where

$$\tau_{\boldsymbol{u},j}(n) = \log n + \log(1 \vee R_j) + \log\left(1 \vee \|u_j\|_{\mathbf{x}}\right),$$

and where $C > 0$ depends only on k and $\alpha_1, \ldots, \alpha_k$.

Roughly speaking, this result says that the risk bound we get when $\sqrt{s}$ is of the form $\sqrt{s(t,x)} = f_{\boldsymbol{u}(x)}(t)$, corresponds to the one we would get if we could apply a model selection theorem on the components $u_1, \ldots, u_k$ separately. Each term $\varepsilon_j(u_j)$ can be controlled under structural or smoothness assumptions on u_j. For instance, if $\mathbb{X} = [0,1]^{k_2}$ and if u_j is assumed to belong to the class $\mathscr{F}_j = \mathcal{H}([0,1]^{k_2})$, a suitable choice of $(\mathbb{W}_j, \Delta_j)$ leads to

$$C_j \varepsilon_j(u_j) \leq (R_j L(u_j)^{\alpha_j})^{\frac{2k_2}{k_2 + 2\alpha_j \boldsymbol{\beta}_j}} \left(\frac{\tau_{\boldsymbol{u},j}(n)}{n} \right)^{\frac{2\alpha_j \bar{\boldsymbol{\beta}}_j}{2\alpha_j \boldsymbol{\beta}_j + k_2}} + \frac{\tau_{\boldsymbol{u},j}(n)}{n}$$

where $\boldsymbol{\beta}_j$ is such that $u_j \in \mathcal{H}^{\boldsymbol{\beta}_j}([0,1]^{k_2})$ and where $C_j > 0$ depends only on k_2 and $\boldsymbol{\beta}_j$. In particular, if $\alpha_j = 1$ and if n is large, $\varepsilon_j(u_j)$ is of order $(\log n/n)^{2\bar{\boldsymbol{\beta}}_j/(2\bar{\boldsymbol{\beta}}_j + k_2)}$. Apart from the logarithmic factor, this corresponds to the estimation rate of an Hölderian function on $[0,1]^{k_2}$.

The corollary below illustrates this result for the Duane model.

Corollary 4. *There exists an estimator $\hat{s}$ such that, for all $\boldsymbol{\alpha} \in (0,+\infty)^{k_2}$, $\boldsymbol{\beta} \in (0,+\infty)^{k_2}$, for all $a \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_2})$, $b \in \mathcal{H}^{\boldsymbol{\beta}}([0,1]^{k_2})$ satisfying $b > -1/2$, and for all function f of the form $f(t,x) = a(x)t^{b(x)}$,*

$$\begin{aligned} C\mathbb{E}\left[H^2(s,\hat{s})\right] \ \leq \ & d_2^2\left(\sqrt{s}, f\right) \\ & + \left(\frac{1}{1 \wedge \inf_{x \in [0,1]^{k_2}} (2b(x)+1)} \right)^{\frac{k_2}{2\bar{\boldsymbol{\alpha}} + k_2}} L(a)^{\frac{2k_2}{2\bar{\boldsymbol{\alpha}} + k_2}} \left(\frac{\log n}{n} \right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}} + k_2}} \\ & + \left(\frac{1 \vee \|a\|_\infty^2}{1 \wedge \inf_{x \in [0,1]^{k_2}} (2b(x)+1)^3} \right)^{\frac{k_2}{2\bar{\boldsymbol{\beta}} + k_2}} L(b)^{\frac{2k_2}{2\bar{\boldsymbol{\beta}} + k_2}} \left(\frac{\log n}{n} \right)^{\frac{2\bar{\boldsymbol{\beta}}}{2\bar{\boldsymbol{\beta}} + k_2}} \\ & + C' \frac{\log n}{n} \end{aligned}$$

where $C > 0$ depends on k_2, $\max_{1 \leq j \leq k} \alpha_j$, $\max_{1 \leq j \leq k} \beta_j$, and where C' depends on $L(a)$, $L(b)$, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\beta}}$, $\|a\|_\infty$, $\|b\|_\infty$ and $\inf_{x \in [0,1]^{k_2}} (2b(x)+1)$.

5.2. Change point detection. In the case where the intensity s_i of each N_i is of the form $\sqrt{s_i(t)} = f_{\boldsymbol{\theta}_i}(t)$, a natural way to control the risk of our estimator $\hat{s}$ is to consider some assumptions on the map $i \mapsto \boldsymbol{\theta}_i$. This problem amounts to choosing suitable collections $\mathbb{W}_1, \ldots, \mathbb{W}_k$ to approximate functions on $\mathbb{X} = \{1, \ldots, n\}$.

In this section, we focus on the case where the map $i \mapsto \boldsymbol{\theta}_i$ is piecewise constant with a small number of jumps. Let $\mathcal{P}$ be the set of partitions of $\{1, \dots, n\}$ into intervals. We aim at estimating s when there exists a partition $P_0 \in \mathcal{P}$ such that s is of the form

$$\forall I \in P_0, \exists \boldsymbol{\theta}_I \in \Theta, \forall i \in I, \quad \sqrt{s_i(t)} = f_{\boldsymbol{\theta}_I}(t) \quad \text{for all } t \in \mathbb{T}. \tag{12}$$

We define for each partition $P \in \mathcal{P}$, the linear space of piecewise constant functions

$$W_P = \left\{ \sum_{I \in P} a_I \mathbb{1}_I,\ a_I \in \mathbb{R} \right\}$$

and apply Theorem 6 with the collections and maps defined by

$$\forall j \in \{1, \dots, k\},\ \mathbb{W}_j = \{W_P,\ P \in \mathcal{P}\} \text{ and } \Delta_j(W_P) = |P| + \log \binom{n-1}{|P|-1}.$$

This leads to the result below.

Corollary 5. *Assume that Assumption 2 and relation (12) hold. There exists an estimator $\hat{s}$ such that*

$$C\mathbb{E}\left[H^2(s, \hat{s})\right] \leq |P_0| \frac{\log n + C'}{n},$$

where $C > 0$ depends only on k and $\alpha_1, \dots, \alpha_k$, where C' is given by

$$C' = \sup_{1 \leq j \leq k} \left(\log\left(1 + R_j\right)\right) + \sup_{I \in P} \left(\log\left(1 + \|\boldsymbol{\theta}_I\|_\infty\right)\right),$$

and where $\|\boldsymbol{\theta}_I\|_\infty = \sup_{1 \leq j \leq k} |(\boldsymbol{\theta}_I)_j|$.

For illustration purpose, in the context of the Duane model, there exist $a_1 \dots, a_n \in (0, +\infty)$, and $b_1, \dots, b_n \in (-1/2, +\infty)$ such that $\sqrt{s_i(t)} = a_i t^{b_i}$ for all $t \in (0, 1]$. By combining the preceding corollary with Proposition 2, we build an estimator $\hat{s}$ such that

$$C\mathbb{E}\left[H^2(s, \hat{s})\right] \leq (1 + r_1 + r_2) \frac{\log n + C'}{n}$$

where r_1 and r_2 are the numbers of jumps of the maps $i \mapsto a_i$ and $i \mapsto b_i$ respectively, where C is an universal positive constant, and where C' depends on $\sup_{1 \leq i \leq n} a_i$, $\sup_{1 \leq i \leq n} |b_i|$ and $\inf_{1 \leq i \leq n}(2b_i + 1)$.

The preceding collections $\mathbb{W}_1, \dots, \mathbb{W}_k$ can also be used to approximate the map $i \mapsto \boldsymbol{\theta}_i$ under other assumptions such as smoothness ones. For instance, an approximation theorem for monotone functions on $\{1, \dots, n\}$ can be found in Baraud and Birgé (2009) and can be used to deal with the situation where some components of the map $i \mapsto \boldsymbol{\theta}_i$ are monotone.

6. Proofs

6.1. Proof of Theorem 1. Throughout the proof, we set $\mathbf{N} = (N_1, \dots, N_n)$ and $\mathbf{x} = (x_1, \dots, x_n)$.

6.1.1. About the T-estimators. We begin to briefly recall the general strategy introduced in Birgé (2006) to build estimators from tests.

Given two distinct functions f, f' of $\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$, a test function $\psi_{f,f'}(\mathbf{N},\mathbf{x})$ is a measurable function with values in $\{f, f'\}$. The convention is that $\psi_{f,f'}(\mathbf{N},\mathbf{x}) = f$ means accepting f whereas $\psi_{f,f'}(\mathbf{N},\mathbf{x}) = f'$ means accepting f'. In what follows, we need tests with the following properties. We shall build them in Section 6.1.2.

Assumption 3. *There exist $a > 0$, $\kappa > 0$ such that for all distinct functions $f, f' \in \mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$ and all $z \in \mathbb{R}$, there exists a test $\psi_{f,f'}^{(z)}(\mathbf{N},\mathbf{x})$ satisfying*

$$\sup_{\substack{f\in\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X},M),\\ \kappa H(s,f)\le H(f,f')}} \mathbb{P}\left[\psi_{f,f'}^{(z)}(\mathbf{N},\mathbf{x}) = f'\right] \le \exp\left[-an\left(H^2(f,f') + z\right)\right] \tag{13}$$

$$\sup_{\substack{f\in\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X},M),\\ \kappa H(s,f')\le H(f,f')}} \mathbb{P}\left[\psi_{f,f'}^{(z)}(\mathbf{N},\mathbf{x}) = f\right] \le \exp\left[-an\left(H^2(f,f') - z\right)\right]. \tag{14}$$

We now consider an at most countable collection $\mathbb{S}$ of subsets of $\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$. We shall assume that the sets $S \in \mathbb{S}$ are D-models. We recall the definition below.

Definition 2. *A subset S of $\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$ is called a D-model with parameters $\bar\eta_S$, $\bar D_S$ and 1 if*

$$|S \cap \mathcal{B}(f, x\bar\eta_S)| \le \exp\left[\bar D_S x^2\right] \quad \text{for all } x \ge 2 \text{ and } f \in \mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M),$$

where $\mathcal{B}(f, x\bar\eta_S)$ is the closed ball centered at f with radius $x\bar\eta_S$ of the metric space $(\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M), H)$.

The tests allow to select among the functions of $\cup_{S\in\mathbb{S}} S$. Precisely, the selection rule is the following.

Given a collection $\mathbb{S}$ of D-models, we set for all $f \in \cup_{S\in\mathbb{S}} S$,

$$\bar\eta(f) = \inf\{\bar\eta_S,\ S \in \mathbb{S},\ S \ni f\}$$

and for all $f' \in \cup_{S\in\mathbb{S}} S$, $f' \ne f$, $z_{f,f'} = \bar\eta(f')^2 - \bar\eta(f)^2$. We define for all $f \in \cup_{S\in\mathbb{S}} S$,

$$\mathcal{R}(f) = \left\{f' \in \cup_{S\in\mathbb{S}} S,\ \psi_{f,f'}^{(z_{f,f'})}(\mathbf{N},\mathbf{x}) = f'\right\}$$

and consider

$$\gamma(f) = \begin{cases} \sup\{H(f,f'),\ f' \in \mathcal{R}(f)\} & \text{if } \mathcal{R}(f) \ne \emptyset,\\ 0 & \text{if } \mathcal{R}(f) = \emptyset. \end{cases}$$

Given $\varepsilon > 0$, a T_ε-estimator is a measurable function $\hat s = \hat s(\mathbf{N},\mathbf{x})$ with values in $\cup_{S\in\mathbb{S}} S$ such that

$$\gamma(\hat s) \vee \varepsilon\bar\eta(\hat s) = \inf_{f\in\cup_{S\in\mathbb{S}} S}\left[\gamma(f) \vee \varepsilon\bar\eta(f)\right].$$

Theorem 5 of Birgé (2006) shows that such a minimizer exists almost surely and they all possess similar theoretical properties. In our framework, we can rewrite it as follows.

Theorem 7. *Suppose that Assumption 3 holds. Let $\mathbb{S}$ be an at most countable collection of D-models such that $\bar{D}_S \geq 1/2$ for all $S \in \mathbb{S}$,*

$$\sum_{S\in\mathbb{S}} \exp\left(-\frac{an\bar{\eta}_S^2}{21}\right) \leq 1 \quad \text{and} \quad an\bar{\eta}_S^2 \geq \frac{21\bar{D}_S}{5} \quad \text{for all } S \in \mathbb{S}.$$

For all $\varepsilon \in (0,4]$, there exists almost surely a T_ε-estimator $\hat{s} \in \mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$. Moreover, any of them satisfies

$$\mathbb{P}\left[CH^2(s,\hat{s}) \geq \inf_{S\in\mathbb{S}}\left\{H^2(s,S) + \bar{\eta}_S^2\right\} + \xi\right] \leq e^{-n\xi} \quad \text{for all } \xi > 0,$$

where $C > 0$ depends only on a, κ.

It remains thus to construct the tests and the collection $\mathbb{S}$ to prove Theorem 1.

6.1.2. Definition of the tests Our tests are inspired from the variational formula in Baraud (2011). Let, for all functions f, f' of $\mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$, $T_{f,f'}(\mathbf{N},\mathbf{x})$ be the functional

$$\begin{aligned} T_{f,f'}(\mathbf{N},\mathbf{x}) &= \frac{1}{2n}\sum_{i=1}^n \int_{\mathbb{T}} \sqrt{\frac{f(t,x_i)+f'(t,x_i)}{2}}\left(\sqrt{f'(t,x_i)} - \sqrt{f(t,x_i)}\right) \mathrm{d}\mu(t) \\ &\quad + \frac{1}{\sqrt{2}n}\sum_{i=1}^n \int_{\mathbb{T}} \frac{\sqrt{f'(t,x_i)} - \sqrt{f(t,x_i)}}{\sqrt{f(t,x_i)+f'(t,x_i)}} \mathrm{d}N_i(t) \\ &\quad - \frac{1}{2n}\sum_{i=1}^n \int_{\mathbb{T}} \left(f'(t,x_i) - f(t,x_i)\right) \mathrm{d}\mu(t) \end{aligned}$$

where the convention $0/0$ is in use. We prove the following.

Lemma 6. *There exist positive numbers a, b such that for all $z \in \mathbb{R}$ and all $f, f' \in \mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$ satisfying $4H(s,f) \leq H(f,f')$,*

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N},\mathbf{x}) \geq bz\right] \leq \exp\left[-na\left(H^2(f,f') + z\right)\right].$$

The proof of this lemma is delayed to Section 6.1.6 and we refer to the proof for the exact values of a and b.

This lemma says that the functional $T_{f,f'}(\mathbf{N},\mathbf{x})$ can be used to construct the tests. Precisely, we set for all $z \in \mathbb{R}$, $f, f' \in \mathbb{L}_+^1(\mathbb{T}\times\mathbb{X}, M)$, $f \neq f'$,

$$\psi_{f,f'}^{(z)}(\mathbf{N},\mathbf{x}) = \begin{cases} f' & \text{if } T_{f,f'}(\mathbf{N},\mathbf{x}) > bz \\ f & \text{if } T_{f,f'}(\mathbf{N},\mathbf{x}) < bz, \end{cases}$$

and $\psi_{f,f'}^{(z)}(\mathbf{N},\mathbf{x})$ is defined arbitrary in case of equality. Thanks to the above lemma, (13) holds. Note that (14) also holds since $T_{f,f'}(\mathbf{N},\mathbf{x}) = -T_{f',f}(\mathbf{N},\mathbf{x})$.

6.1.3. Construction of $\mathbb{S}$ The collection $\mathbb{S}$ is derived from $\mathbb{V}$. We shall show in Section 6.1.6 the following lemma.

Lemma 7. *For all $\eta > 0$ and $V \in \mathbb{V}$ there exists a D-model $\bar{S}_V(\eta)$ with parameters η, $63D_V(\eta/2)$ and 1. Moreover,*

$$H(s, \bar{S}_V(\eta)) \leq 2\sqrt{2}\left(d_2\left(\sqrt{s}, V\right) + \eta\right) \tag{15}$$

and for all $f \in \bar{S}_V(\eta)$, there exists $g \in V$ such that $\sqrt{f} = g \vee 0$.

Please note that we can assume (for the sake of simplicity and with no loss of generality), that D_V is non-increasing. We then set for all $V \in \mathbb{V}$,

$$\bar{\eta}_{\bar{S}_V} = \left(21\sqrt{\frac{3}{5a}}\eta_V\right) \vee \sqrt{\frac{21\Delta(V)}{na}} \quad \text{and} \quad \bar{S}_V = \bar{S}_V(\bar{\eta}_{\bar{S}_V})$$

where a is given by Lemma 6. Actually a is very small (smaller than 1), which implies that $\bar{\eta}_{\bar{S}_V}/2 \geq \eta_V$ and thus

$$63D_V(\bar{\eta}_{\bar{S}_V}/2) \leq 63D_V(\eta_V).$$

Consequently, the set $\bar{S}_V$ is a D-model with parameters $\bar{\eta}_{\bar{S}_V}$, $\bar{D}_{\bar{S}_V} = 63D_V(\eta_V)$ and 1. The collection $\mathbb{S}$ is then defined by $\mathbb{S} = \left\{\bar{S}_V,\, V \in \mathbb{V}\right\}$.

6.1.4. Proof of Theorem 1. The assumptions of Theorem 7 are fulfilled:

$$an\bar{\eta}^2_{\bar{S}_V} \geq \frac{21^2 \times 3}{5} n\eta_V^2 \geq \frac{21^2 \times 3}{5} D_V(\eta_V) \geq \frac{21\bar{D}_{\bar{S}_V}}{5}$$

and

$$\sum_{V \in \mathbb{V}} \exp\left(-\frac{an\bar{\eta}^2_{\bar{S}_V}}{21}\right) \leq \sum_{V \in \mathbb{V}} \exp\left(-\Delta(V)\right) \leq 1.$$

The selection rule described in Section 6.1.1 provides thus an estimator $\hat{s} \in \cup_{V \in \mathbb{V}} \bar{S}_V$ such that, for all $\xi > 0$,

$$\mathbb{P}\left[CH^2(s, \hat{s}) \geq \inf_{V \in \mathbb{V}} \left\{H^2\left(s, \bar{S}_V\right) + \bar{\eta}^2_{\bar{S}_V}\right\} + \xi\right] \leq e^{-n\xi}$$

where $C > 0$ is universal. By using inequality (15),

$$H^2\left(s, \bar{S}_V\right) \leq 16\left[d_2^2(\sqrt{s}, V) + \bar{\eta}^2_{\bar{S}_V}\right]$$

and hence

$$\inf_{V \in \mathbb{V}} \left\{H^2\left(s, \bar{S}_V\right) + \bar{\eta}^2_{\bar{S}_V}\right\} \leq C' \inf_{V \in \mathbb{V}} \left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\}$$

for some universal constant $C' > 0$. Finally,

$$\mathbb{P}\left[C''H^2(s, \hat{s}) \geq \inf_{V \in \mathbb{V}} \left\{d_2^2\left(\sqrt{s}, V\right) + \eta_V^2 + \frac{\Delta(V)}{n}\right\} + \xi\right] \leq e^{-n\xi}$$

where $C'' = C/(C' \vee 1)$. □

6.1.5. Proof of Lemma 6. We start with the following Bennett-type inequality which generalizes Proposition 7 of Reynaud-Bouret (2003).

Lemma 8. *Let $f_1, \dots, f_n$ be n bounded measurable functions. Let ρ, υ be positive numbers such that $\rho \geq \max_{1\leq i\leq n} \|f_i\|_\infty$ and*

$$\frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{T}} f_i^2(t) s_i(t)\, \mathrm{d}\mu(t) \leq \upsilon.$$

Then, for all $r \geq 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\int_{\mathbb{T}} f_i(t)\,\mathrm{d}N_i(t) - \mathbb{E}\left(\int_{\mathbb{T}} f_i(t)\,\mathrm{d}N_i(t)\right)\right] \geq r\right) &\leq \exp\left(-n\frac{\upsilon}{b^2} h\left(\frac{\rho r}{\upsilon}\right)\right)\\ &\leq \exp\left(-n\frac{r^2}{2\left(\upsilon + \frac{\rho r}{3}\right)}\right)\end{aligned}$$

where h is the function defined for $u \in (-1, +\infty)$ by $h(u) = (1+u)\log(1+u) - u$.

Proof. By homogeneity we can assume that $\rho = 1$. We assume moreover that for each $i \in \{1, \dots, n\}$, f_i is a piecewise constant function (with a finite number of pieces). There exist thus $k_1, \dots, k_n \in \mathbb{N}^\star$ and a family $(a_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq k_i}}$ of elements of $[-1, 1]$ such that

$$\forall t \in \mathbb{T}, \quad f_i(t) = \sum_{j=1}^{k_i} a_{i,j} \mathbb{1}_{A_{i,j}}(t)$$

where the $A_{i,j}$ are measurable sets of $\mathbb{T}$ such that $A_{i,j} \cap A_{i,j'} = \emptyset$ for all $j \neq j'$.

Now, for all $\xi > 0$,

$$\begin{aligned}\log \mathbb{E}\left(e^{\xi \sum_{i=1}^n \left[\int_{\mathbb{T}} f_i\,\mathrm{d}N_i - \mathbb{E}\left(\int_{\mathbb{T}} f_i\,\mathrm{d}N_i\right)\right]}\right) &= \sum_{i=1}^{n} \log \mathbb{E}\left(e^{\xi\left[\int_{\mathbb{T}} f_i\,\mathrm{d}N_i - \mathbb{E}\left(\int_{\mathbb{T}} f_i\,\mathrm{d}N_i\right)\right]}\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k_i} \log \mathbb{E}\left(e^{\xi a_{i,j}\left[N_i(A_{i,j}) - \mathbb{E}(N_i(A_{i,j}))\right]}\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k_i} \mathbb{E}\left(N_i(A_{i,j})\right)\left(e^{\xi a_{i,j}} - \xi a_{i,j} - 1\right).\end{aligned}$$

By using the monotony of the function $x \mapsto (e^x - x - 1)/x^2$,

$$\begin{aligned}\log \mathbb{E}\left(e^{\xi \sum_{i=1}^n \left[\int_{\mathbb{T}} f_i\,\mathrm{d}N_i - \mathbb{E}\left(\int_{\mathbb{T}} f_i\,\mathrm{d}N_i\right)\right]}\right) &\leq \sum_{i=1}^{n}\sum_{j=1}^{k_i} \mathbb{E}\left(a_{i,j}^2 N_i(A_{i,j})\right)\left(e^{\xi} - \xi - 1\right)\\ &\leq n\upsilon(e^{\xi} - \xi - 1).\end{aligned}$$

This inequality still holds when the f_i are not piecewise constant since a measurable function can be approximated by piecewise constant functions. Indeed, there exists a sequence $(f_i^{(k)})_{k\geq 1}$ of

piecewise constant functions (with a finite number of jumps) such that $f_i^{(k)} \to f_i$ when $k \to +\infty$ in the space $\mathbb{L}^2(\mathbb{T}, s_i \, \mathrm{d}\mu)$ and such that $\|f_i^{(k)}\|_\infty \leq 1$ whatever k, i. By using Fatou lemma,

$$\log \mathbb{E}\left(\liminf_{k \to +\infty} e^{\xi \sum_{i=1}^n \left[\int_{\mathbb{T}} f_i^{(k)} \, \mathrm{d}N_i - \int_{\mathbb{T}} f_i^{(k)} s_i \, \mathrm{d}\mu_i \right]} \right) \leq n\upsilon(e^\xi - \xi - 1).$$

Since,

$$\mathbb{E}\left[\left| \int_{\mathbb{T}} f_i^{(k)} \, \mathrm{d}N_i - \int_{\mathbb{T}} f_i \, \mathrm{d}N_i \right| \right] \leq \int_{\mathbb{T}} \left| f_i^{(k)} - f_i \right| s_i \, \mathrm{d}\mu \to 0$$

one can assume (up to considering a subsequence) that $\int_{\mathbb{T}} f_i^{(k)} \, \mathrm{d}N_i - \int_{\mathbb{T}} f_i \, \mathrm{d}N_i \to 0$ almost surely (for all $i \in \{1, \dots n\}$). We then have

$$\log \mathbb{E}\left(e^{\xi \sum_{i=1}^n \left[\int_{\mathbb{T}} f_i \, \mathrm{d}N_i - \int_{\mathbb{T}} f_i s_i \, \mathrm{d}\mu_i \right]} \right) \leq n\upsilon(e^\xi - \xi - 1)$$

as wished. The exponential inequality is then deduced from the Cramér-Chernoff method, see Chapter 2 of Massart (2003). □

Let us return to the proof of Lemma 6. We define the function ζ on $[0, +\infty)^2$ by

$$\zeta(x, y) = \frac{1}{\sqrt{2}} \left(\sqrt{\frac{y}{x+y}} - \sqrt{\frac{x}{x+y}} \right) \quad \text{for all } x, y \in [0, +\infty),$$

where we use the convention $0/0 = 0$. Let then

$$\begin{aligned} Z_{f,f'}(\mathbf{N}, \mathbf{x}) &= T_{f,f'}(\mathbf{N}, \mathbf{x}) - \mathbb{E}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \right] \\ &= \int_{\mathbb{T} \times \mathbb{X}} \zeta(f, f') \, \mathrm{d}M - \mathbb{E}\left(\int_{\mathbb{T} \times \mathbb{X}} \zeta(f, f') \, \mathrm{d}M \right). \end{aligned}$$

We use the claim below whose proof ensues from the proofs of Propositions 2 and 3 of Baraud (2011).

Claim 1.

$$\mathbb{E}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \right] \leq \left(1 + \frac{1}{\sqrt{2}} \right) H^2(s, f) - \left(1 - \frac{1}{\sqrt{2}} \right) H^2(s, f')$$

and

$$\frac{1}{n} \sum_{i=1}^n \int_{\mathbb{T}} \zeta^2 \left(f(t, x_i), f'(t, x_i) \right) s(t, x_i) \, \mathrm{d}\mu(t) \leq H^2(s, f) + H^2(s, f') + H^2(f, f').$$

We derive from the first point of the claim that

$$\begin{aligned} \mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z \right] &= \mathbb{P}\left[Z_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z - \mathbb{E}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \right] \right] \\ &\leq \mathbb{P}\left[Z_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z - \left(1 + \frac{1}{\sqrt{2}} \right) H^2(s, f) + \left(1 - \frac{1}{\sqrt{2}} \right) H^2(s, f') \right]. \end{aligned}$$

Note that the random variable $Z(f, f')$ can be written as

$$Z_{f,f'}(\mathbf{N}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \left[\int_{\mathbb{T}} \zeta\left(f(\cdot, x_i), f'(\cdot, x_i) \right) \, \mathrm{d}N_i - \mathbb{E}\left(\int_{\mathbb{T}} \zeta\left(f(\cdot, x_i), f'(\cdot, x_i) \right) \, \mathrm{d}N_i \right) \right].$$

When

$$r = z - \left(1 + \frac{1}{\sqrt{2}}\right) H^2(s, f) + \left(1 - \frac{1}{\sqrt{2}}\right) H^2(s, f')$$

is non-negative, we apply Lemma 8 with $f_i(\cdot) = \zeta\left(f(\cdot, x_i), f'(\cdot, x_i)\right)$, $\rho = 1/\sqrt{2}$ and

$$\upsilon = H^2(s, f) + H^2(s, f') + H^2(f, f')$$

to obtain

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z\right] \leq \exp\left(-\frac{nr^2}{2\upsilon + \frac{r\sqrt{2}}{3}}\right).$$

We now bound from above the right-hand side of this inequality.

For this, we begin to bound υ from above. We deduce from the triangular inequality and from $4H(s, f) \leq H(f, f')$ that

$$\begin{aligned} \upsilon &\leq 3H^2(s, f) + 3H^2(f, f') \\ &\leq 3(1 + 1/16)H^2(f, f'). \end{aligned}$$

Now, we bound r from below. Note that

$$H(f, f') \leq H(s, f) + H(s, f') \leq \frac{1}{4}H(f, f') + H(s, f')$$

and thus $H(s, f') \geq 3/4H(f, f')$. This leads to

$$\begin{aligned} r &\geq z - \left(1 + \frac{1}{\sqrt{2}}\right)\frac{1}{16}H^2(f, f') + \left(1 - \frac{1}{\sqrt{2}}\right)\frac{9}{16}H^2(f, f') \\ &\geq z + CH^2(f, f') \end{aligned}$$

where $C = (8 - 5\sqrt{2})/16 > 0$. There are two types of cases involved.

- If $z + CH^2(f, f') > 0$, r is non-negative and thus

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z\right] \leq \exp\left(-\frac{n\left(z + CH^2(f, f')\right)^2}{6(1 + 1/16)H^2(f, f') + \frac{\sqrt{2}}{3}\left(z + CH^2(f, f')\right)}\right).$$

  Set $C' = 9\sqrt{2}(1 + 1/16) + C$. Then,

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z\right] \leq \exp\left(-\frac{3n}{\sqrt{2}}\frac{\left(z + CH^2(f, f')\right)^2}{z + C'H^2(f, f')}\right).$$

  One can then verify that

$$\begin{aligned} \frac{(z + CH^2(f, f'))^2}{z + C'H^2(f, f')} &= \frac{C^2}{C'}H^2(f, f') + \frac{(2C' - C)C}{C'^2}z + \frac{(C - C')^2 z^2}{C'^2(z + C'H^2(f, f'))} \\ &\geq \frac{C^2}{C'}H^2(f, f') + \frac{(2C' - C)C}{C'^2}z \end{aligned}$$

  which implies that

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq z\right] \leq \exp\left(-\frac{3n}{\sqrt{2}}\left(\frac{C^2}{C'}H^2(f, f') + \frac{(2C' - C)C}{C'^2}z\right)\right). \tag{16}$$

- If now $z + CH^2(f, f') \leq 0$,

$$\begin{aligned} \frac{C^2}{C'} H^2(f, f') + \frac{(2C' - C)C}{C'^2} z &\leq \left(\frac{C^2}{C'} - \frac{(2C' - C)C^2}{C'^2} \right) H^2(f, f') \\ &\leq 0. \end{aligned}$$

  Consequently, (16) also holds.

We thus have proved that

$$\mathbb{P}\left[T_{f,f'}(\mathbf{N}, \mathbf{x}) \geq bz\right] \leq \exp\left[-na\left(H^2(f, f') + z\right)\right]$$

where

$$\begin{aligned} a &= \frac{3C^2}{\sqrt{2}C'} \simeq 4.5 \times 10^{-4} \\ b &= \frac{CC'}{2C' - C} \simeq 0.029. \end{aligned}$$

This ends the proof. □

6.1.6. Proof of Lemma 7. By using Proposition 7 of Birgé (2006), we derive from $S_V(\eta)$ a set $S'_V(\eta) \subset V$ such that

$$\forall \varphi \in \mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M),\ \forall x \geq 2, \quad |S'_V(\eta) \cap \mathcal{B}(\varphi, x\eta)| \leq \exp\left(7D_V(\eta/2)x^2\right)$$

where $\mathcal{B}(\varphi, x\eta)$ is the ball centered at φ with radius $x\eta$ of the metric space $(\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M), d_2)$, and such that

$$\forall f \in V, \quad d_2(f, S'_V(\eta)) \leq \eta.$$

Proposition 12 of Birgé (2006) (applied with $T = S'_V(\eta)$, the cone M_0 of non-negative functions of $\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M)$, $M' = \mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M)$ and $\bar{\pi}$ defined by $\bar{\pi}(f) = f \vee 0$) provides a subset $S''_V(\eta)$ such that the functions $f \in S''_V(\eta)$ are non negative, such that

$$\forall f \in \mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M),\ \forall x \geq 2, \quad |S''_V(\eta) \cap \mathcal{B}(f, x\eta)| \leq \exp\left(63D_V(\eta)\, x^2\right)$$

and such that

$$\text{for all non-negative function } f \in \mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M), \quad d_2(f, S''_V(\eta)) \leq 4d_2(f, S'_V(\eta)).$$

The lemma holds with $\bar{S}_V(\eta) = \{\sqrt{f},\ f \in S''_V(\eta)\}$. □

6.2. Proof of Lemma 1. The proof of this proposition requires the following elementary lemma.

Lemma 9. *Let $f, f' \in \mathbb{L}^2(\mathbb{T}, \mu)$ and $g, g' \in \mathbb{L}^2(\mathbb{X}, \nu_n)$ such that $\|f\|_{\mathtt{t}} = \|f'\|_{\mathtt{t}} = 1$ and $\|g\|_{\mathtt{x}} = \|g'\|_{\mathtt{x}} = 1$. Let $\kappa, \kappa' \in \mathbb{R}$. Then,*

$$d_2^2\left(\kappa f g, \kappa' f' g'\right) = \left(\kappa - \kappa'\right)^2 + \kappa\kappa'\left(d_{\mathtt{t}}^2(f, f') + d_{\mathtt{x}}^2(g, g') - 1/2\, d_{\mathtt{t}}^2(f, f') d_{\mathtt{x}}^2(g, g')\right).$$

Let $\eta > 0$. In this proof, we say that a set $S(\eta)$ is a η-net of a set V in a metric space (E, d) if, for all $y \in V$, there exists $x \in S(\eta)$ such that $d(x, y) \leq \eta$.

Let us denote by S_1 (respectively S_2) the unit sphere of V_1 (respectively V_2). Let $S_1(\eta) \subset S_1$ (respectively $S_2(\eta) \subset S_2$) be a η-net of S_1 (respectively S_2) such that

$$\forall f \in V_1,\ \forall x \geq 0, \quad |S_1(\eta) \cap \mathcal{B}_t(f, x\eta)| \leq (2x+1)^{\dim V_1} \tag{17}$$

$$\forall g \in V_2,\ \forall x \geq 0, \quad |S_2(\eta) \cap \mathcal{B}_{\mathtt{x}}(g, x\eta)| \leq (2x+1)^{\dim V_2} \tag{18}$$

where $\mathcal{B}_t(f, x\eta)$ and $\mathcal{B}_{\mathtt{x}}(g, x\eta)$ are the closed balls centered at f and g with radius $x\eta$ of the metric spaces $(\mathbb{L}^2(\mathbb{T}, \mu), d_{\mathtt{t}})$ and $(\mathbb{L}^2(\mathbb{X}, \nu_n), d_{\mathtt{x}})$ respectively. We refer to Lemma 4 of Birgé (2006) for the existence of these nets. Let now

$$S(\eta) = \bigcup_{k \in \mathbb{N}^\star} \left\{ \frac{k\eta}{\sqrt{2}} fg,\ (f, g) \in S_1\left(\frac{1}{\sqrt{2}k}\right) \times S_2\left(\frac{1}{\sqrt{2}k}\right) \right\}.$$

First of all, $S(\eta)$ is a η-net of V. Indeed, let $\varphi \in V$. We can write $\varphi(t, x) = \kappa f(t) g(x)$ where $\kappa \geq 0$, $f \in S_1$ and $g \in S_2$. Let us define

$$k = \inf\left\{ i \in \mathbb{N}^\star,\ i \geq \sqrt{2}\kappa/\eta \right\}$$

and let $(f', g') \in S_1(1/(\sqrt{2}k)) \times S_2(1/(\sqrt{2}k))$ such that

$$d_{\mathtt{t}}(f, f') \leq \frac{1}{\sqrt{2}k} \quad \text{and} \quad d_{\mathtt{x}}(g, g') \leq \frac{1}{\sqrt{2}k}.$$

By using Lemma 9, the application $\varphi'(t, x) = \frac{k\eta}{\sqrt{2}} f'(t) g'(x)$ is such that

$$d_2\left(\varphi, \varphi'\right) \leq \eta,$$

which proves that $S(\eta)$ is a η-net of V.

According to Definition 1, we now consider $\varphi \in \mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M)$ and $x \geq 2$ and aim at bounding from above the cardinality of the set $S(\eta) \cap \mathcal{B}\left(\varphi, x\eta\right)$ (where we recall that $\mathcal{B}\left(\varphi, x\eta\right)$ is the closed ball centered at φ with radius $x\eta$ of the metric space $(\mathbb{L}^2(\mathbb{T} \times \mathbb{X}, M), d_2)$).

For this purpose, we begin to assume that φ belongs to $S(\eta)$, which implies that the function can be written as $\varphi(t, x) = \kappa f(t) g(x)$. We introduce

$$\mathcal{K} = \left\{ \frac{k\eta}{\sqrt{2}},\ k \in \mathbb{N}^\star,\ \left| \frac{k\eta}{\sqrt{2}} - \kappa \right| \leq x\eta \right\}$$

and for all $\kappa' \in \mathcal{K}$,

$$\mathcal{C}(\kappa') = \left(S_1\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{t}}\left(f, 6x^2 \frac{\eta}{2\kappa'}\right) \right) \times \left(S_2\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{x}}\left(g, 6x^2 \frac{\eta}{2\kappa'}\right) \right).$$

Let

$$T(\eta) = \left\{ \kappa' f' g',\ \kappa' \in \mathcal{K},\ (f', g') \in \mathcal{C}(\kappa') \right\}.$$

We shall prove that

$$S(\eta) \cap \mathcal{B}\left(\kappa f g, x\eta\right) \subset T(\eta). \tag{19}$$

We then upper-bound the cardinality of $S(\eta) \cap \mathcal{B}(\kappa f g, x\eta)$ by bounding from above the cardinality of $T(\eta)$.

Let $\varphi' \in S(\eta) \cap \mathcal{B}(\kappa f g, x\eta)$. There exist κ', f' and g' such that $\varphi' = \kappa' f' g'$ and we derive from Lemma 9 that

$$(\kappa - \kappa')^2 \leq d_2^2(\varphi, \varphi') \leq x^2 \eta^2,$$

which implies that $\kappa' \in \mathcal{K}$. We now distinguish several cases.

- Suppose that
$$\left(\int_{\mathbb{T}} f(t) f'(t) \,\mathrm{d}\mu(t)\right)\left(\int_{\mathbb{X}} g(x) g'(x) \,\mathrm{d}\nu_n(x)\right) < 0.$$
We then have $d_2^2(\varphi, \varphi') \geq \kappa^2 + \kappa'^2$. Since $\kappa \geq \eta/\sqrt{2}$, $\kappa' \leq \kappa + x\eta$ and $x \geq 2$,
$$\frac{\kappa'}{\kappa} \leq 1 + \sqrt{2}x \leq \frac{3}{2}x.$$
Thus, $d_2^2(\varphi, \varphi') \geq 4\kappa'^2/(9x^2)$. Since $f, f' \in S_1$ and $g, g' \in S_2$,
$$\|f - f'\|_{\mathtt{t}}^2 \leq 4 \quad \text{and} \quad \|g - g'\|_{\mathtt{x}}^2 \leq 4$$
and thus
$$\|f - f'\|_{\mathtt{t}}^2 \leq \frac{9x^2}{\kappa'^2} d_2^2(\varphi, \varphi') \leq \frac{9x^4}{\kappa'^2}\eta^2$$
$$\|g - g'\|_{\mathtt{x}}^2 \leq \frac{9x^2}{\kappa'^2} d_2^2(\varphi, \varphi') \leq \frac{9x^4}{\kappa'^2}\eta^2.$$
We then have
$$f' \in S_1\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{t}}\left(f, 6x^2\frac{\eta}{2\kappa'}\right) \quad \text{and} \quad g' \in S_2\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{x}}\left(g, 6x^2\frac{\eta}{2\kappa'}\right)$$
that is $(f', g') \in \mathcal{C}(\kappa')$ and thus $\varphi' \in T(\eta)$.

- If now,
$$\int_{\mathbb{T}} f(t) f'(t) \,\mathrm{d}\mu(t) > 0 \quad \text{and} \quad \int_{\mathbb{X}} g(x) g'(x) \,\mathrm{d}\nu_n(x) > 0,$$
then $d_{\mathtt{t}}^2(f, f') \leq 2$ and $d_{\mathtt{x}}^2(g, g') \leq 2$. We then derive from Lemma 9 and from the elementary inequality
$$\frac{1}{2}(y_1 + y_2) \leq y_1 + y_2 - \frac{1}{2}y_1 y_2 \quad \text{for all } y_1, y_2 \in [0, 2],$$
that
$$(\kappa - \kappa')^2 + \frac{\kappa\kappa'}{2}\left(d_{\mathtt{t}}^2(f, f') + d_{\mathtt{x}}^2(g, g')\right) \leq d_2^2(\varphi, \varphi') \leq x^2\eta^2.$$
Hence,
$$d_{\mathtt{t}}^2(f, f') + d_{\mathtt{x}}^2(g, g') \leq \frac{2x^2\eta^2}{\kappa\kappa'}.$$
By using the inequality $\kappa'/\kappa \leq 3/2x$ proved in the first point, we deduce
$$f' \in S_1\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{t}}\left(f, 2\sqrt{3}x^{3/2}\frac{\eta}{2\kappa'}\right) \quad \text{and} \quad g' \in S_2\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathtt{x}}\left(g, 2\sqrt{3}x^{3/2}\frac{\eta}{2\kappa'}\right).$$
Since $2\sqrt{3}x^{3/2} \leq 6x^2$ (because $x \geq 2$), we have $(f', g') \in \mathcal{C}(\kappa')$ and thus $\varphi' \in T(\eta)$.

- Finally, assume that

$$\int_{\mathbb{T}} f(t)f'(t)\,\mathrm{d}\mu(t) < 0 \quad \text{and} \quad \int_{\mathbb{X}} g(x)g'(x)\,\mathrm{d}\nu_n(x) < 0.$$

  Note that the function φ' can also be written as $\varphi' = \kappa'(-f')(-g')$. We then deduce from the second point that

$$-f' \in S_1\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathfrak{t}}\left(f, 2\sqrt{3}x^{3/2}\frac{\eta}{2\kappa'}\right) \text{ and } -g' \in S_2\left(\frac{\eta}{2\kappa'}\right) \cap \mathcal{B}_{\mathfrak{x}}\left(g, 2\sqrt{3}x^{3/2}\frac{\eta}{2\kappa'}\right)$$

  and thus $(-f', -g') \in \mathcal{C}(\kappa')$. Hence, $\varphi' \in T(\eta)$ as wished.

We thus have proved (19) and thus

$$|S(\eta) \cap \mathcal{B}(\kappa f g, x\eta)| \leq \sum_{\kappa' \in \mathcal{K}} \left|\mathcal{C}(\kappa')\right|.$$

Now, note that $|\mathcal{K}| \leq 2\sqrt{2}x + 1$. By using (17) and (18), for all κ',

$$|\mathcal{C}(\kappa')| \leq \left(12x^2+1\right)^{\dim V_1 + \dim V_2}.$$

Consequently, we have proved

$$\forall \varphi \in S(\eta),\, \forall x \geq 2, \quad |S(\eta) \cap \mathcal{B}(\varphi, x\eta)| \leq \left(2\sqrt{2}x+1\right)\left(12x^2+1\right)^{\dim V_1 + \dim V_2}.$$

Let us recall that we must to upper bound the cardinality of $S(\eta) \cap \mathcal{B}(\varphi, x\eta)$ for all $\varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$. For this, if $\varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$, may be $|S(\eta) \cap \mathcal{B}(\varphi, x\eta)| = 0$. If not, there exists $\varphi' \in S(\eta) \cap \mathcal{B}(\varphi, x\eta)$ and thus

$$|S(\eta) \cap \mathcal{B}(\varphi, x\eta)| \leq \left|S(\eta) \cap \mathcal{B}\left(\varphi', 2x\eta\right)\right|.$$

Consequently, for all $\varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$,

$$\forall x \geq 2, \quad |S(\eta) \cap \mathcal{B}(\varphi, x\eta)| \leq \left(4\sqrt{2}x+1\right)\left(48x^2+1\right)^{\dim V_1 + \dim V_2}.$$

The conclusion ensues from the elementary inequalities

$$\forall x \geq 2, \quad 4\sqrt{2}x + 1 \leq e^{1.4x^2} \text{ and } 48x^2 + 1 \leq e^{1.4x^2}.$$

□

6.3. Proof of Proposition 3. For all pair $(V_1, V_2) \in \mathbb{V}_1 \times \mathbb{V}_2$, we define the set V by relation (6). Let then $\mathbb{V}$ be the collection of all V when (V_1, V_2) varies among $\mathbb{V}_1 \times \mathbb{V}_2$. Let $\bar{\Delta}$ be the application on $\mathbb{V}$ defined by

$$\bar{\Delta}(V) = \Delta_1(V_1) + \Delta_2(V_2)$$

when V corresponds to (V_1, V_2). We apply Theorem 1 with $\mathbb{V}$ and $\bar{\Delta}$ to derive an estimator $\hat{s}$ such that

$$C\mathbb{E}\left[H^2(s,\hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ d_2^2\left(\sqrt{s}, V\right) + \frac{\dim V_1 + \dim V_2 + 1 + \Delta_1(V_1) + \Delta_2(V_2)}{n} \right\}.$$

Let $\kappa v_1 v_2 \in \mathscr{F}$, and let $(v_1', v_2') \in V_1 \times V_2$ such that $\|v_1'\|_{\mathtt{t}} = \|v_2'\|_{\mathtt{x}} = 1$. The preceding inequality implies

$$\begin{aligned} C'\mathbb{E}\left[H^2(s,\hat{s})\right] \quad \leq \quad & d_2^2\left(\sqrt{s}, \kappa v_1 v_2\right) + \kappa^2 d_2^2\left(v_1 v_2, v_1' v_2'\right) \\ & + \frac{\dim V_1 \vee 1 + \dim V_2 \vee 1 + \Delta_1(V_1) + \Delta_2(V_2)}{n} \end{aligned}$$

where $C' = C/2$. By using Lemma 9 (page 45),

$$d_2^2\left(v_1 v_2, v_1' v_2'\right) \leq d_{\mathtt{t}}^2\left(v_1, v_1'\right) + d_{\mathtt{x}}^2\left(v_2, v_2'\right).$$

By taking the infimum over all v_1' and v_2',

$$\begin{aligned} C'\mathbb{E}\left[H^2(s,\hat{s})\right] \quad \leq \quad & d_2^2\left(\sqrt{s}, \kappa v_1 v_2\right) + \kappa^2 d_{\mathtt{t}}^2\left(v_1, S_1\right) + \frac{\dim V_1 \vee 1 + \Delta_1(V_1)}{n} \\ & + \kappa^2 d_{\mathtt{x}}^2\left(v_2, S_2\right) + \frac{\dim V_2 \vee 1 + \Delta_2(V_2)}{n} \end{aligned}$$

where S_1 and S_2 are the unit spheres of V_1 and V_2 respectively.

Now, remark that

$$d_{\mathtt{t}}(v_1, S_1) \leq 2 d_{\mathtt{t}}(v_1, V_1) \quad \text{and} \quad d_{\mathtt{x}}(v_2, S_2) \leq 2 d_{\mathtt{x}}(v_2, V_2).$$

Indeed, if w_1 is the projection of v_1 on V_1, then

$$\begin{aligned} d_{\mathtt{t}}(v_1, S_1) \quad & \leq \quad \left\| v_1 - \frac{w_1}{\|w_1\|_{\mathtt{t}}} \right\|_{\mathtt{t}} \\ & \leq \quad \|v_1 - w_1\|_{\mathtt{t}} + \left\| w_1 - \frac{w_1}{\|w_1\|_{\mathtt{t}}} \right\|_{\mathtt{t}}. \end{aligned}$$

Now,

$$\left\| w_1 - \frac{w_1}{\|w_1\|_{\mathtt{t}}} \right\|_{\mathtt{t}} = \left| 1 - \frac{1}{\|w_1\|_{\mathtt{t}}} \right| \|w_1\|_{\mathtt{t}} = \left| \|w_1\|_{\mathtt{t}} - 1 \right|.$$

Since $\|v_1\|_{\mathtt{t}} = 1$,

$$\left\| w_1 - \frac{w_1}{\|w_1\|_{\mathtt{t}}} \right\|_{\mathtt{t}} = \left| \|w_1\|_{\mathtt{t}} - \|v_1\|_{\mathtt{t}} \right| \leq \|v_1 - w_1\|_{\mathtt{t}}.$$

Since $\|v_1 - w_1\|_{\mathtt{t}} = d_{\mathtt{t}}(v_1, V_1)$, we have $d_{\mathtt{t}}(v_1, S_1) \leq 2 d_{\mathtt{t}}(v_1, V_1)$. The proof of the inequality $d_{\mathtt{x}}(v_2, S_2) \leq 2 d_{\mathtt{x}}(v_2, V_2)$ is similar.

The conclusion follows. □

6.4. Proof of Lemma 2. For all $b, b' \in (-1/2, +\infty)$,

$$\int_0^1 \left(\sqrt{2b+1}t^b - \sqrt{2b'+1}t^{b'}\right)^2 \mathrm{d}t = \frac{4\left(b-b'\right)^2}{\left(1+b+b'\right)\left(\sqrt{2b+1}+\sqrt{2b'+1}\right)^2}$$

and thus

$$\frac{\left(b-b'\right)^2}{(1+2(b\vee b'))^2} \leq \int_0^1 \left(\sqrt{2b+1}t^b - \sqrt{2b'+1}t^{b'}\right)^2 \mathrm{d}t \leq \frac{\left(b-b'\right)^2}{(1+2(b\wedge b'))^2}$$

which ends the proof. □

6.5. Proof of Lemma 3. For $b > 0$, we define

$$g_b(t) = \frac{f_b(t)}{\|f_b\|_{\mathfrak{t}}} = \frac{2^{k/2}b^{1/2+k/2}}{\sqrt{k/2\,k!}}\, t^{k/2}e^{-bt}.$$

For all $b, b' > 0$,

$$\begin{aligned}
\frac{1}{2}\int_0^\infty \left(g_b(t) - g_{b'}(t)\right)^2 \mathrm{d}t &= 1 - \frac{\left(2\sqrt{bb'}\right)^{k+1}}{(b+b')^{k+1}} \\
&= \frac{\sum_{j=0}^k (b+b')^{k-j}(2\sqrt{bb'})^j}{(b+b')^{k+1}(\sqrt{b}+\sqrt{b'})^2}\left(b-b'\right)^2 \\
&= \frac{1}{(b+b')(\sqrt{b}+\sqrt{b'})^2}\sum_{j=0}^k \left(\frac{2\sqrt{bb'}}{b+b'}\right)^j \left(b-b'\right)^2.
\end{aligned}$$

Consequently,

$$\frac{1}{8(b\vee b')^2}\left(b-b'\right)^2 \leq \frac{1}{2}\int_0^\infty \left(g_b(t) - g_{b'}(t)\right)^2 \mathrm{d}t \leq \frac{k+1}{8(b\wedge b')^2}\left(b-b'\right)^2$$

which concludes the proof. □

6.6. Proof of Proposition 5. We generalize Lemma 1 for some new spaces. The proof of the following lemma is analogous to the one of Lemma 1 and will not be detailed.

Lemma 10. *Let V_1 and V_2 be subsets of the unit spheres of $\mathbb{L}^2(\mathbb{T}, \mu)$ and $\mathbb{L}^2(\mathbb{X}, \nu_n)$ respectively. For each $i \in \{1, 2\}$, we assume that there exist positive numbers $\underline{\rho_i}, \bar{\rho}_i$, a subset $W_{\mathtt{i}}$ of a finite dimensional normed linear space $(\bar{W}_{\mathtt{i}}, |\cdot|_{\mathtt{i}})$ and a surjective map Φ_i from $W_{\mathtt{i}}$ onto V_i such that:*

$$\forall (x,y) \in W_1, \quad \underline{\rho_1}|x-y|_1 \leq d_{\mathfrak{t}}(\Phi_1(x), \Phi_1(y)) \leq \bar{\rho}_1|x-y|_1 \tag{20}$$

$$\forall (x,y) \in W_2, \quad \underline{\rho_2}|x-y|_2 \leq d_{\mathfrak{x}}(\Phi_2(x), \Phi_2(y)) \leq \bar{\rho}_2|x-y|_2. \tag{21}$$

The set

$$V = \{\kappa v_1 v_2,\ (v_1, v_2) \in V_1 \times V_2,\ \kappa \in [0, +\infty)\}$$

has a finite metric dimension bounded by

$$D_V = C\left[1+\log\left(1+\frac{\bar{\rho}_1}{\underline{\rho_1}}\right)\dim \bar{W}_1 + \log\left(1+\frac{\bar{\rho}_2}{\underline{\rho_2}}\right)\dim \bar{W}_2\right]$$

where C is an universal constant.

Lemma 11. *Let for all $r, R \in (b_0, +\infty)$, such that $R > r$, $V_{\mathtt{t}}(r, R)$ be the set defined by*

$$V_{\mathtt{t}}(r,R) = \left\{\frac{u_b}{\|u_b\|_{\mathtt{t}}},\, b\in[r,R]\right\}.$$

Condition (20) holds with $\dim \bar{W}_1 = 1$, $\underline{\rho_1} = \underline{\rho}(R)$ and $\bar{\rho_1} = \bar{\rho}(r)$.

Lemma 12. *For all positive number ρ and $W \in \mathbb{W}$, let*

$$V_2(W,\rho) = \left\{\frac{v_{\boldsymbol{\theta}}}{\|v_{\boldsymbol{\theta}}\|_{\mathtt{x}}},\, \boldsymbol{\theta}\in W,\, \|\boldsymbol{\theta}\|\le\rho\right\}.$$

There exists a finite dimensional normed linear space $(\bar{W}_2, |\cdot|_2)$ and a map Φ_2 from $\bar{W}_2$ onto $V_2(W,\rho)$ such that condition (21) holds with $\dim \bar{W}_2 \le \dim W$, $\underline{\rho_2} = e^{-6\rho}$ and $\bar{\rho}_2 = e^{6\rho}$.

Proof of Lemma 12. For any integers $i, j \in \mathbb{N}^\star$, let us denote by $\varphi_{i,j}$ the linear form on $\mathbb{R}^{k_2}$ defined by $\varphi_{i,j}(\boldsymbol{\theta}) = < x_i - x_j, \boldsymbol{\theta} >$ where $< \cdot,\cdot >$ is the standard scalar product on $\mathbb{R}^{k_2}$. Let $W_1 = \cap_{i\neq j}\mathrm{Ker}\,\varphi_{i,j}$ and let W_2 such that $W = W_1 \oplus W_2$ and such that $< u, v >= 0$ for all $(u,v) \in W_1 \times W_2$. Since the functions of $\mathbb{L}^2(\mathbb{X}, \nu_n)$ are defined ν_n-almost everywhere, the set $V_2(W,\rho)$ can be written as

$$V_2(W,\rho) = \Phi_2\left(\{\boldsymbol{\theta}\in W_2,\, \|\boldsymbol{\theta}\|\le\rho\}\right) \quad \text{where} \quad \Phi_2(\boldsymbol{\theta}) = \frac{v_{\boldsymbol{\theta}}}{\|v_{\boldsymbol{\theta}}\|_{\mathtt{x}}}.$$

Indeed, let $\boldsymbol{\theta} \in W$ written as $\boldsymbol{\theta} = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2$ where $\boldsymbol{\theta}_1 \in W_1$ and $\boldsymbol{\theta}_2 \in W_2$. Then, for all $j \in \{1,\dots,n\}$,

$$\begin{aligned}
\frac{v_{\boldsymbol{\theta}}(x_j)}{\|v_{\boldsymbol{\theta}}\|_{\mathtt{x}}} &= \frac{\exp\left(< x_j, \boldsymbol{\theta} >\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \exp\left(2 < x_i, \boldsymbol{\theta} >\right)}} \\
&= \frac{\exp\left(< x_j, \boldsymbol{\theta}_1 > + < x_j, \boldsymbol{\theta}_2 >\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \exp\left(2 < x_i, \boldsymbol{\theta}_1 > + 2 < x_j, \boldsymbol{\theta}_2 >\right)}} \\
&= \frac{v_{\boldsymbol{\theta}_1}(x_j)}{\|v_{\boldsymbol{\theta}_1}\|_{\mathtt{x}}}
\end{aligned}$$

and thus $\Phi_2(\boldsymbol{\theta}) = \Phi_2(\boldsymbol{\theta}_1)$, ν_n-almost everywhere.

For all $x \in \mathbb{X}$, let Ψ_x be the function defined from $\mathbb{X}$ into $\mathbb{R}$ by $\Psi_x(\boldsymbol{\theta}) = \Phi_2(\boldsymbol{\theta})(x) = v_{\boldsymbol{\theta}}(x)/\|v_{\boldsymbol{\theta}}\|_{\mathtt{x}}$. We derive from some calculus that the differential of Ψ_x at the point $\boldsymbol{\theta} \in W_2$, denoted by $d\Psi_x(\boldsymbol{\theta})$, is

$$\forall h \in \mathbb{R}^{k_2}, \quad d\Psi_x(\boldsymbol{\theta})\cdot h = \frac{\frac{1}{n}\sum_{i=1}^n \exp\left(2 < \boldsymbol{\theta}, x_i > + < \boldsymbol{\theta}, x >\right)\left(< x - x_i, h >\right)}{\left(\frac{1}{n}\sum_{i=1}^n \exp\left(2 < \boldsymbol{\theta}, x_i >\right)\right)^{3/2}}.$$

In particular, we have

$$\forall h \in \mathbb{R}^{k_2}, \quad \frac{e^{-6\rho}}{n}\sum_{i=1}^{n} |< x - x_i, h >| \leq |d\Psi_x(\boldsymbol{\theta}) \cdot h| \leq \frac{e^{6\rho}}{n}\sum_{i=1}^{n} |< x - x_i, h >|.$$

If we endow W_2 with the norm $|\cdot|_2$ defined by

$$\forall \boldsymbol{\theta} \in W_2, \quad |\boldsymbol{\theta}|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}|< x_i - x_j, \boldsymbol{\theta} >|\right)^2},$$

the mean value theorem leads to

$$\forall (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in W_2, \quad e^{-6\rho}|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_2 \leq d_{\mathbf{x}}\left(\Phi_2(\boldsymbol{\theta}_1), \Phi_2(\boldsymbol{\theta}_2)\right) \leq e^{6\rho}|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_2,$$

which concludes the proof. □

We now prove Proposition 5. We derive from Lemma 10 that for all $\varrho \geq 1$, all $r, R > b_0$ and all $W \in \mathbb{W}$, the set

$$V(r, R, W, \varrho) = \left\{a u_b v_{\boldsymbol{\theta}},\, a \in [0, +\infty),\, b \in [r, R],\, \boldsymbol{\theta}' \in W,\, \|\boldsymbol{\theta}'\| \leq \varrho\right\}$$

has a metric dimension bounded by

$$CD_{V(r,R,W,\varrho)} = 1 + \varrho \dim W + \log\left(1 + \frac{\bar{\rho}(r)}{\underline{\rho}(R)}\right)$$

for some universal positive constant C.

Let us define the collection $\mathbb{V}$ by

$$\mathbb{V} = \left\{V\left(b_0 + 1/r, b_0 + R, W, \varrho\right),\, W \in \mathbb{W},\, r, R, \varrho \in \mathbb{N}^\star\right\}$$

and the map $\bar{\Delta}$ on $\mathbb{V}$ by

$$\bar{\Delta}\left(V(r, R, W, \varrho)\right) = \Delta(W) + \log(2R^2) + \log(2r^2) + \log(2\varrho^2).$$

We apply Theorem 1 with $(\mathbb{V}, \bar{\Delta})$ to build an estimator $\hat{s}$. For all $W \in \mathbb{W}$, $\varrho, r, R \in \mathbb{N}^\star$, $\boldsymbol{\theta}' \in W$ such that $\|\boldsymbol{\theta}'\| \leq \varrho$, $a \in [0, +\infty)$, $b \in [b_0 + 1/r, b_0 + R]$, this estimator satisfies

$$\begin{aligned} C'\mathbb{E}\left[H^2(s, \hat{s})\right] \quad \leq \quad & d_2^2(\sqrt{s}, a u_b v_{\boldsymbol{\theta}'}) \\ & + \frac{1 + \varrho \dim W + \log\left(1 + \frac{\bar{\rho}(b_0 + 1/r)}{\underline{\rho}(b_0 + R)}\right) + \Delta(W) + \log r + \log R + \log \varrho}{n} \end{aligned}$$

where C' is another universal positive constant.

We may roughly upper-bound the right-hand side of this inequality to get

$$\begin{aligned} C''\mathbb{E}\left[H^2(s, \hat{s})\right] \quad \leq \quad & d_2^2(\sqrt{s}, a u_b v_{\boldsymbol{\theta}'}) \\ & + \frac{1 + \varrho \dim W + \log\left(1 \vee \bar{\rho}\left(b_0 + 1/r\right)\right) + \left|\log\left(1 \wedge \underline{\rho}(b_0 + R)\right)\right| + \Delta(W)}{n} \\ & + \frac{\log r + \log R + \log \varrho}{n} \end{aligned}$$

for some universal positive constant C''.

In particular, for all $W \in \mathbb{W}$, $\boldsymbol{\theta}' \in W$, $a \in [0,+\infty)$ and $b \in I$, we may use this inequality with $R = \inf\{i \in \mathbb{N}^\star,\, i \geq b - b_0\}$, $r = \inf\{i \in \mathbb{N}^\star, i \geq 1/(b-b_0)\}$, $\varrho = \inf\{i \in \mathbb{N}^\star, i \geq \|\boldsymbol{\theta}'\|\}$ to derive

$$\begin{aligned} C'''\mathbb{E}\left[H^2(s,\hat{s})\right] &\leq d_2^2(\sqrt{s}, au_b v_{\boldsymbol{\theta}'}) + \frac{(1 \vee \|\boldsymbol{\theta}'\|)(1 \vee \dim W) + \Delta(W)}{n} \\ &\quad + \frac{1}{n}\left\{\log\left[1 \vee \bar{\rho}\left(b_0 + \frac{b-b_0}{b-b_0+1}\right)\right] + \left|\log\left(1 \wedge \underline{\rho}(1+b)\right)\right| + |\log(b-b_0)|\right\} \end{aligned}$$

where C''' is an universal positive constant.

Now, by using the triangular inequality, we have for all $\boldsymbol{\theta} \in \mathbb{R}^{k_2}$,

$$\begin{aligned} d_2^2(\sqrt{s}, au_b v_{\boldsymbol{\theta}'}) &\leq 2\left(d_2^2(\sqrt{s}, au_b v_{\boldsymbol{\theta}}) + d_2^2(au_b v_{\boldsymbol{\theta}}, au_b v_{\boldsymbol{\theta}'}\right) \\ &\leq 2\left(d_2^2(\sqrt{s}, au_b v_{\boldsymbol{\theta}}) + a^2\|u_b\|_{\mathtt{t}}^2 d_{\mathtt{x}}^2(v_{\boldsymbol{\theta}}, v_{\boldsymbol{\theta}'})\right). \end{aligned}$$

Some calculus shows that $d_{\mathtt{x}}(v_{\boldsymbol{\theta}}, v_{\boldsymbol{\theta}'}) \leq e^{\|\boldsymbol{\theta}\| \vee \|\boldsymbol{\theta}'\|}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

Consequently, for all $a \in [0,+\infty)$, $b \in I$, $\boldsymbol{\theta} \in \mathbb{R}^k$, $W \in \mathbb{W}$, we obtain (by taking $\boldsymbol{\theta}'$ the projection of $\boldsymbol{\theta}$ on W),

$$\begin{aligned} C''''\mathbb{E}\left[H^2(s,\hat{s})\right] &\leq d_2^2(\sqrt{s}, au_b v_{\boldsymbol{\theta}}) + a^2\|u_b\|_{\mathtt{t}}^2 e^{2\|\theta\|} d^2(\boldsymbol{\theta}, W) + \frac{(1 \vee \dim W)(1 \vee \|\boldsymbol{\theta}\|) + \Delta(W)}{n} \\ &\quad + \frac{1}{n}\left\{\log\left[1 \vee \bar{\rho}\left(b_0 + \frac{b-b_0}{b-b_0+1}\right)\right] + \left|\log\left(1 \wedge \underline{\rho}(1+b)\right)\right| + |\log(b-b_0)|\right\} \end{aligned}$$

where C'''' is an universal positive constant. We conclude by taking the infimum over all $W \in \mathbb{W}$. □

6.7. Proofs of Lemmas 4 and 5.

Proof of Lemma 4. We derive from some calculus that for all $\theta_2, \theta_2' \in [-1/2 + 1/r_2, +\infty)$,

$$\int_0^1 (t^{\theta_2} - t^{\theta_2'})^2 \,\mathrm{d}t = \frac{2(\theta_2 - \theta_2')^2}{(1+2\theta_2)(1+\theta_2+\theta_2')(1+2\theta_2')} \leq 2r_2^3(\theta_2 - \theta_2')^2.$$

Hence, for all $(\theta_1, \theta_2), (\theta_1', \theta_2') \in [-r_1, r_1] \times [-1/2 + 1/r_2, +\infty)$,

$$\begin{aligned} \sqrt{\int_0^1 (\theta_1 t^{\theta_2} - \theta_1' t^{\theta_2'})^2 \,\mathrm{d}t} &\leq \sqrt{\int_0^1 (\theta_1 - \theta_1')^2 t^{2\theta_2'} \,\mathrm{d}t} + \sqrt{\int_0^1 \theta_1^2 (t^{\theta_2} - t^{\theta_2'})^2 \,\mathrm{d}t} \\ &\leq \frac{|\theta_1 - \theta_1'|}{\sqrt{2\theta_2' + 1}} + \frac{\sqrt{2}|\theta_1||\theta_2 - \theta_2'|}{\sqrt{(1+2\theta_2)(1+\theta_2+\theta_2')(1+2\theta_2')}} \\ &\leq r_2^{1/2}|\theta_1 - \theta_1'| + \sqrt{2} r_1 r_2^{3/2}|\theta_2 - \theta_2'|. \end{aligned}$$

This ends the proof. □

Proof of Lemma 5. We derive from some calculus that for all $\theta_2, \theta_2' \geq 1/r_2$,

$$\int_0^\infty \left(t^{k/2}e^{-\theta_2 t} - t^{k/2}e^{-\theta_2' t}\right)^2 \mathrm{d}t = \frac{k!}{2^{k+1}}\left(\frac{1}{\theta_2^k} + \frac{1}{\theta_2'^k} - \frac{2^{2+k}}{(\theta_2+\theta_2')^k}\right).$$

If $k = 0$,

$$\int_0^\infty \left(t^{k/2}e^{-\theta_2 t} - t^{k/2}e^{-\theta_2' t}\right)^2 \mathrm{d}t = \frac{(\theta_2'-\theta_2)^2}{2\theta_2'^2\theta_2 + 2\theta_2'\theta_2^2} \leq \frac{r_2^3}{4}(\theta_2'-\theta_2)^2$$

while if $k = 1$,

$$\begin{aligned}\int_0^\infty \left(t^{k/2}e^{-\theta_2 t} - t^{k/2}e^{-\theta_2' t}\right)^2 \mathrm{d}t &= \frac{\theta_2'^2 + 4\theta_2'\theta_2 + \theta_2^2}{4\theta_2'^2\theta_2^2(\theta_2'+\theta_2)^2}(\theta_2'-\theta_2)^2 \\ &\leq \frac{3r_2^4}{8}(\theta_2'-\theta_2)^2.\end{aligned}$$

Hence, for all $(\theta_1,\theta_2),(\theta_1',\theta_2') \in [-r_1,r_1]\times[1/r_2,+\infty)$,

$$\begin{aligned}\sqrt{\int_0^\infty \left(\theta_1 t^{k/2}e^{-\theta_2 t} - \theta_1' t^{k/2}e^{-\theta_2' t}\right)^2 \mathrm{d}t} &\leq \sqrt{\int_0^\infty (\theta_1-\theta_1')^2 t^k e^{-2\theta_2' t}\,\mathrm{d}t} \\ &\quad + |\theta_1|\sqrt{\int_0^\infty \left(t^{k/2}e^{-\theta_2 t} - t^{k/2}e^{-\theta_2' t}\right)^2 \mathrm{d}t}\end{aligned}$$

Now,

$$\sqrt{\int_0^\infty (\theta_1-\theta_1')^2 t^k e^{-2\theta_2' t}\,\mathrm{d}t} = \frac{\sqrt{k!}|\theta_1-\theta_1'|}{2^{(k+1)/2}\theta_2'^{(k+1)/2}},$$

which ends the proof. □

6.8. Proof of Theorem 6. We start with the following proposition.

Proposition 8. *Suppose that Assumption 2 holds. Let for all $j \in \{1,\dots,k\}$, W_j be a linear subspace of $\mathbb{L}^2(\mathbb{X},\nu_n)$ with finite dimension and Z_j be a bounded subset of W_j. Let then $\boldsymbol{\rho} \in [0,+\infty)^k$ such that for all $j \in \{1,\dots,k\}$, $Z_j \subset \mathcal{B}_{\mathbf{x}}(0,\rho_j) = \{g \in \mathbb{L}^2(\mathbb{X},\nu_n),\ \|g\|_{\mathbf{x}} \leq \rho_j\}$. Let $m_1,\dots,m_k \in \mathbb{R}\cup\{-\infty\}$ and $M_1,\dots,M_k \in \mathbb{R}\cup\{\infty\}$ be such that*

$$\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^k,\ \forall i \in \{1,\dots,k\},\ m_i \leq x_i \leq M_i\right\}$$

and let π be the map defined on $\mathbb{R}^k$ by

$$\pi(\boldsymbol{x}) = \big((x_1 \vee m_1) \wedge M_1, \dots, (x_k \vee m_k) \wedge M_k\big) \quad \textit{for all } \boldsymbol{x} = (x_1,\dots,x_k) \in \mathbb{R}^k.$$

Let for all $\boldsymbol{u} \in \prod_{j=1}^k Z_j$, $g_{\boldsymbol{u}}$ be the function defined by

$$g_{\boldsymbol{u}(x)}(t) = f_{\pi(\boldsymbol{u}(x))}(t) \quad \textit{for all } (t,x) \in \mathbb{T}\times\mathbb{X}.$$

Then, the set V defined by

$$V = \left\{ g_{\boldsymbol{u}},\, \boldsymbol{u} \in \prod_{j=1}^{k} Z_j \right\}$$

has a metric dimension bounded by

$$D_V(\eta) = \frac{1}{2} \vee \frac{1}{4} \sum_{j=1}^{k} \log \left(1 + 2 \left(\frac{kR_j}{\eta} \right)^{1/\alpha_j} \rho_j \right) \dim(W_j).$$

Proof of Proposition 8. As in the proof of Lemma 1, we say that a set $S(\eta)$ is a η-net of a set V in a metric space (E, d) if, for all $y \in V$, there exists $x \in S(\eta)$ such that $d(x, y) \leq \eta$.

Let $\eta > 0$ and for $j \in \{1, \ldots, k\}$,

$$\eta_j = \left(\frac{\eta}{kR_j} \right)^{1/\alpha_j}.$$

Let $Z_j'(\eta_j)$ be a maximal subset of Z_j such that $d_{\mathbf{x}}(x, y) > \eta_j$ for all $x \neq y \in Z_j'(\eta)$. This is a η_j-net of Z_j such that

$$|Z_j'(\eta_j)| \leq \left| Z_j'(\eta_j) \cap \mathcal{B}_{\mathbf{x}}(0, \rho_j) \right|$$

and by using Lemma 4 of Birgé (2006),

$$|Z_j'(\eta_j)| \leq \left(\frac{2\rho_j}{\eta_j} + 1 \right)^{\dim W_j}. \tag{22}$$

To prove the proposition, we begin to show that the set

$$S(\eta) = \left\{ g_{\boldsymbol{u}},\, \boldsymbol{u} \in \prod_{j=1}^{k} Z_j'\left(\eta_j\right) \right\}$$

is a η-net of V.

Let $f \in V$ be the function of the form $f(t, x) = g_{\boldsymbol{u}(x)}(t) = f_{\pi(\boldsymbol{u}(x))}(t)$ and for all $j \in \{1, \ldots, k\}$, let $v_j \in Z_j'\left(\eta_j\right)$ such that $d_{\mathbf{x}}(u_j, v_j) \leq \eta_j$. We define $\boldsymbol{v} = (v_1, \ldots, v_k)$ and $g \in S(\eta)$ by $g(t, x) = g_{\boldsymbol{v}(x)}(t) = f_{\pi(\boldsymbol{v}(x))}(t)$. Then,

$$\begin{aligned} \|f - g\|_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \left\| f_{\pi(\boldsymbol{u}(x_i))}(\cdot) - f_{\pi(\boldsymbol{v}(x_i))}(\cdot) \right\|_{\mathbf{t}}^2 \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_j \left| u_j(x_i) - v_j(x_i) \right|^{\alpha_j} \right)^2. \end{aligned}$$

By using the Cauchy-Schwarz inequality,

$$\begin{aligned} \|f - g\|_2^2 &\leq \frac{1}{n} \sum_{i=1}^{n} k \left(\sum_{j=1}^{k} R_j^2 \left| u_j(x_i) - v_j(x_i) \right|^{2\alpha_j} \right) \\ &\leq k \sum_{j=1}^{k} R_j^2 \left(\frac{1}{n} \sum_{i=1}^{n} \left| u_j(x_i) - v_j(x_i) \right|^{2\alpha_j} \right). \end{aligned}$$

By using the concavity of the map $x \mapsto x^{\alpha_j}$,

$$\begin{aligned}\|f-g\|_2^2 &\leq k\sum_{j=1}^{k} R_j^2 \left(\frac{1}{n}\sum_{i=1}^{n}\left|u_j(x_i)-v_j(x_i)\right|^2\right)^{\alpha_j} \\ &\leq k\sum_{j=1}^{k} R_j^2 d_{\mathbf{x}}^{2\alpha_j}(u_j,v_j) \\ &\leq \eta^2\end{aligned}$$

as wished.

We now consider $x \geq 2$, $\varphi \in \mathbb{L}^2(\mathbb{T}\times\mathbb{X}, M)$ and aim at bounding from above the cardinality of $S(\eta)\cap\mathcal{B}(\varphi, x\eta)$. We have,

$$|S(\eta)\cap\mathcal{B}(\varphi,x\eta)| \leq \prod_{j=1}^{k}\left|Z_j'(\eta_j)\right|.$$

By using (22),

$$\begin{aligned}|S(\eta)\cap\mathcal{B}(\varphi,x\eta)| &\leq \prod_{j=1}^{k}\left(2\left(\frac{kR_j}{\eta}\right)^{1/\alpha_j}\rho_j+1\right)^{\dim(W_j)} \\ &\leq \exp\left(\frac{1}{4}\sum_{j=1}^{k}\dim(W_j)\log\left(2\left(\frac{kR_j}{\eta}\right)^{1/\alpha_j}\rho_j+1\right)x^2\right).\end{aligned}$$

This ends the proof. □

Lemma 13. *Let V be a set with metric dimension bounded by D_V. Assume that there exist $k\in\mathbb{N}^\star$, $\boldsymbol{a},\boldsymbol{b}\in[0,+\infty)^k$ such that $\max_{1\leq j\leq k} a_j \geq 1$, $\min_{1\leq j\leq k} b_j \geq 1$ and such that*

$$D_V(\eta) \leq \frac{1}{2}\vee\sum_{j=1}^{k} a_j\log\left(1+\frac{b_j}{\eta}\right) \quad \textit{for all } \eta>0.$$

Then,

$$\eta_V = \inf\left\{\eta>0,\ \frac{D_V(\eta)}{\eta^2}\leq n\right\}$$

can be upper bounded by

$$C\eta_V^2 \leq \frac{\sum_{j=1}^{k} a_j\log(1+b_j)}{n} + \frac{\sum_{j=1}^{k} a_j}{n}\log\left(1+\frac{n}{\sum_{j=1}^{k} a_j}\right)$$

where C is an universal positive constant.

Proof of Lemma 13. For all $\eta>0$,

$$\frac{1}{2}\vee\sum_{j=1}^{k} a_j\log\left(1+\frac{b_j}{\eta}\right) \leq \begin{cases} 2\sum_{j=1}^{k} a_j\log(2b_j) + 2\left(\sum_{j=1}^{k} a_j\right)\log\left(\frac{1}{\eta}\right) & \text{if } \eta<1 \\ 2\sum_{j=1}^{k} a_j\log(2b_j) & \text{otherwise.}\end{cases}$$

The larger D_V, the larger η_V. Consequently, without lost of generality we can assume that

$$D_V(\eta) = \begin{cases} 2\sum_{j=1}^k a_j \log(2b_j) + 2\left(\sum_{j=1}^k a_j\right)\log\left(\frac{1}{\eta}\right) & \text{if } \eta < 1 \\ 2\sum_{j=1}^k a_j \log(2b_j) & \text{otherwise.} \end{cases}$$

Remark now that for all $\alpha, \beta, y > 0$, the equation

$$\alpha + \beta \log x = \frac{y}{2x^2}$$

has only one positive solution x given by

$$x^2 = \frac{y}{\beta L\left(\frac{e^{\frac{2\alpha}{\beta}}}{\beta} y\right)},$$

where L is the Lambert function, defined as being the inverse function of $t \mapsto te^t$. Consequently, by setting

$$\alpha = \sum_{j=1}^k a_j \log(2b_j) \quad \text{and} \quad \beta = \sum_{j=1}^k a_j$$

we derive that the positive number η defined by

$$\eta^2 = \begin{cases} \frac{\beta}{n} L\left(\frac{e^{\frac{2\alpha}{\beta}}}{\beta} n\right) & \text{if } n > 2\alpha \\ \frac{2\alpha}{n} & \text{if } n \le 2\alpha \end{cases}$$

is such that $D_V(\eta) = n\eta^2$. In particular, $\eta_V \le \eta$. The conclusion ensues from some elementary inequalities on the Lambert function. □

We derive from this lemma the following result.

Lemma 14. *Under the notations and assumptions of Proposition 8, there exists an universal positive constant C such that*

$$C\eta_V^2 \le \frac{1}{n}\sum_{j=1}^k \left(\frac{\dim(W_j) \vee 1}{\alpha_j}\right)\left(\log\left(1 + kR_j\rho_j^{\alpha_j}\right) + \log n\right).$$

Proof. We can upper bound $D_V(\eta)$ as follows.

$$\begin{aligned} D_V(\eta) &= \frac{1}{2} \vee \frac{1}{4}\sum_{j=1}^k \log\left(1 + 2\left(\frac{kR_j}{\eta}\right)^{1/\alpha_j}\rho_j\right)\dim(W_j) \\ &\le \frac{1}{2} \vee \frac{1}{4}\sum_{j=1}^k \frac{1}{\alpha_j}\log\left(1 + 2^{\alpha_j}\frac{kR_j}{\eta}\rho_j^{\alpha_j}\right)\dim(W_j). \end{aligned}$$

We then use Lemma 13 with $a_j = \alpha_j^{-1}\left(1 \vee \dim W_j\right)$ and $b_j = 1 \vee \left(2^{\alpha_j} kR_j\rho_j^{\alpha_j}\right)$ (we recall that $\alpha_j \le 1$). There exists thus an universal constant C' such that

$$C'\eta_V^2 \le \frac{1}{n}\sum_{j=1}^k \left(\frac{\dim(W_j) \vee 1}{\alpha_j}\right)\left[\log\left(1 + 1 \vee \left(2^{\alpha_j} kR_j\rho_j^{\alpha_j}\right)\right) + \log\left(1 + \frac{n}{\sum_{i=1}^k \left(\frac{\dim(W_i)\vee 1}{\alpha_i}\right)}\right)\right].$$

We now roughly upper-bound the right-hand side of this inequality to end the proof. □

Let us return to the proof of Theorem 6. For all $W_j \in \mathbb{W}_j$, $\rho_j \in \mathbb{N}^\star$, we introduce the set $Z_j(W_j, \rho_j) = W_j \cap \mathcal{B}_{\mathtt{x}}(0, \rho_j)$ where $\mathcal{B}_{\mathtt{x}}(0, \rho_j)$ is the closed ball centered at 0 with radius ρ_j of the metric space $(\mathbb{L}^2(\mathbb{X}, \nu_n), d_{\mathtt{x}})$. For all $\boldsymbol{W} = (W_1, \dots, W_k) \in \prod_{j=1}^k \mathbb{W}_j$, $\boldsymbol{\rho} = (\rho_1, \dots, \rho_k) \in (\mathbb{N}^\star)^k$, we define

$$V(\boldsymbol{W}, \boldsymbol{\rho}) = \left\{ (t,x) \mapsto f_{\pi(\boldsymbol{u}(x))}(t),\ \boldsymbol{u} \in \prod_{j=1}^k Z_j(W_j, \rho_j) \right\}.$$

We then define

$$\mathbb{V} = \left\{ V(\boldsymbol{W}, \boldsymbol{\rho}),\ \boldsymbol{W} \in \prod_{j=1}^k \mathbb{W}_j,\ \boldsymbol{\rho} \in (\mathbb{N}^\star)^k \right\}$$

and we define the map Δ on $\mathbb{V}$ by

$$\Delta\left(V(\boldsymbol{W}, \boldsymbol{\rho})\right) = \sum_{j=1}^k \left(\Delta_j(W_j) + \log\left(2\rho_j^2\right)\right).$$

We apply Theorem 1 with $(\mathbb{V}, \Delta)$ to build an estimator $\hat{s}$. For all $\boldsymbol{W} = (W_1, \dots, W_k) \in \prod_{j=1}^k \mathbb{W}_j$, $\boldsymbol{\rho} = (\rho_1, \dots, \rho_k) \in (\mathbb{N}^\star)^k$,

$$C\mathbb{E}\left[H^2(s, \hat{s})\right] \leq d_2^2\left(\sqrt{s}, V(\boldsymbol{W}, \boldsymbol{\rho})\right) + \eta^2_{V(\boldsymbol{W}, \boldsymbol{\rho})} + \frac{\Delta\left(V(\boldsymbol{W}, \boldsymbol{\rho})\right)}{n} \tag{23}$$

where C is an universal positive constant. We then derive from Lemma 14 that there exists an universal positive constant C' such that

$$C'\eta^2_{V(\boldsymbol{W}, \boldsymbol{\rho})} \leq \frac{1}{n} \sum_{j=1}^k \left(\frac{\dim(W_j) \vee 1}{\alpha_j}\right) \left(\log\left(1 + kR_j\rho_j^{\alpha_j}\right) + \log n\right).$$

In particular, for all function $f \in \mathscr{F}$ of the form $f(t,x) = f_{\boldsymbol{u}(x)}(t)$, for all $\boldsymbol{W} \in \prod_{j=1}^k \mathbb{W}_j$, for all map $\boldsymbol{v} = (v_1, \dots, v_k) \in \prod_{j=1}^k W_j$, such that for all $j \in \{1, \dots, k\}$, $\|v_j\|_{\mathtt{x}} \leq \|u_j\|_{\mathtt{x}}$, and for all function g of the form $g(t,x) = f_{\pi(\boldsymbol{v}(x))}(t)$, inequality (23) used with $\rho_j = \inf\{i \in \mathbb{N}^\star,\ i \geq \|u_j\|_{\mathtt{x}}\}$ leads to

$$\begin{aligned} C''\mathbb{E}\left[H^2(s, \hat{s})\right] \quad \leq \quad & d_2^2\left(\sqrt{s}, f\right) + d_2^2(f, g) + \frac{\sum_{j=1}^k \left(\Delta_j(W_j) + \log(1 + \|u_j\|_{\mathtt{x}})\right)}{n} \\ & + \frac{1}{n} \sum_{j=1}^k \left(\frac{\dim(W_j) \vee 1}{\alpha_j}\right) \left[\log\left(1 + kR_j(1 + \|u_j\|_{\mathtt{x}})^{\alpha_j}\right) + \log n\right] \end{aligned}$$

where C'' is an universal positive constant.

By using Assumption 2 and the Cauchy-Schwarz inequality,

$$d_2^2(f, g) \leq k \sum_{j=1}^k R_j^2 \|u_j - v_j\|_{\mathtt{x}}^{2\alpha_j}.$$

We then choose v_j as being the projection of u_j on W_j in the space $\mathbb{L}^2(\mathbb{X}, \nu_n)$, and take the infimum over all $\boldsymbol{W} \in \prod_{j=1}^k \mathbb{W}_j$ to conclude.

6.9. Proof of Corollary 4. Let $r_1, r_2 \in \mathbb{N}^\star$. We may apply Theorem 6 with collections $\mathbb{W}_1, \mathbb{W}_2$ provided by Proposition 1 of Baraud and Birgé (2011). This yields an estimator $\tilde{s}$ such that for all $a \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_2})$ with values into $[-r_1, r_1]$, for all $b \in \mathcal{H}^{\boldsymbol{\beta}}([0,1]^{k_2})$ with values into $[-1/2 + 1/r_2, +\infty)$,

$$C\mathbb{E}\left[H^2(s, \tilde{s})\right] \leq d_2^2(\sqrt{s}, f) + \varepsilon_1(a) + \varepsilon_2(b)$$

where

$$\begin{aligned} C_1\varepsilon_1(a) &\leq \left(r_2^{1/2} L(a)\right)^{\frac{2k_2}{k_2+2\bar{\boldsymbol{\alpha}}}} \left(\frac{\log n + \log(1 \vee r_2^{1/2}) + \log\left(1 \vee \|a\|_{\mathtt{x}}\right)}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+k_2}} \\ &\quad + \frac{\log n + \log(1 \vee r_2^{1/2}) + \log\left(1 \vee \|a\|_{\mathtt{x}}\right)}{n} \\ C_2\varepsilon_2(b) &\leq \left(\sqrt{2} r_1 r_2^{3/2} L(b)\right)^{\frac{2k_2}{k_2+2\bar{\boldsymbol{\beta}}}} \left(\frac{\log n + \log(1 \vee \sqrt{2} r_1 r_2^{3/2}) + \log\left(1 \vee \|b\|_{\mathtt{x}}\right)}{n}\right)^{\frac{2\bar{\boldsymbol{\beta}}}{2\bar{\boldsymbol{\beta}}+k_2}} \\ &\quad + \frac{\log n + \log(1 \vee \sqrt{2} r_1 r_2^{3/2}) + \log\left(1 \vee \|b\|_{\mathtt{x}}\right)}{n} \end{aligned}$$

where $C > 0$ is universal, where $C_1 > 0$ depends only on k_2, $\max_{1 \leq j \leq k_2} \alpha_j$, and where $C_2 > 0$ depends only on k_2, $\max_{1 \leq j \leq k_2} \beta_j$.

The above estimator depends on r_1 and r_2. We can thus use Proposition 2, to derive that there exists an estimator $\hat{s}$, such that for all $r_1, r_2 \in \mathbb{N}^\star$, for all $a \in \mathcal{H}^{\boldsymbol{\alpha}}([0,1]^{k_2})$ with values into $[-r_1, r_1]$, for all $b \in \mathcal{H}^{\boldsymbol{\beta}}([0,1]^{k_2})$ with values into $[-1/2 + 1/r_2, +\infty)$,

$$C''\mathbb{E}\left[H^2(s, \hat{s})\right] \leq d_2^2(\sqrt{s}, f) + \varepsilon_1(a) + \varepsilon_2(b) + \frac{\log r_1 + \log r_2}{n}$$

where $C'' > 0$ is universal.

In particular, if we choose r_1 as being the smallest integer larger than $\|a\|_\infty$ and r_2 as being the smallest integer larger than $2/(\inf_{x \in [0,1]^{k_2}}(2b(x) + 1))$, we get

$$C_1'\varepsilon_1(a) \leq \left(\sqrt{1 + \frac{2}{\inf_{x \in [0,1]^{k_2}}(2b(x) + 1)}} L(a)\right)^{\frac{2k_2}{k_2+2\bar{\boldsymbol{\alpha}}}} \left(\frac{\log n}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+k_2}} + C_1'' \frac{\log n}{n}$$

where $C_1' > 0$ depends only on k_2, $\max_{1 \leq j \leq k_2} \alpha_j$ and where C_1'' depends only on k_2, $\bar{\boldsymbol{\alpha}}$, $\|a\|_\infty$, $L(a)$, $\|b\|_\infty$ and $\inf_{x \in [0,1]^{k_2}}(2b(x) + 1)$. We then use

$$1 + \frac{2}{\inf_{x \in [0,1]^{k_2}}(2b(x) + 1)} \leq \frac{3}{1 \wedge \inf_{x \in [0,1]^{k_2}}(2b(x) + 1)}$$

to get

$$C_1'''\varepsilon_1(a) \leq \left(\frac{1}{1 \wedge \inf_{x \in [0,1]^{k_2}}(2b(x) + 1)}\right)^{\frac{k_2}{k_2+2\bar{\boldsymbol{\alpha}}}} L(a)^{\frac{2k_2}{k_2+2\bar{\boldsymbol{\alpha}}}} \left(\frac{\log n}{n}\right)^{\frac{2\bar{\boldsymbol{\alpha}}}{2\bar{\boldsymbol{\alpha}}+k_2}} + C_1'''' \frac{\log n}{n}.$$

We can bound from above $\varepsilon_2(b)$ in a similar fashion. □

References

Antoniadis, A., Besbeas, P., and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā. The Indian Journal of Statistics. Series A*, 63:309–327.

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, 88:805–820.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.

Baraud, Y. and Birgé, L. (2011). Estimating composite functions by model selection. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. To appear.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH.

Comte, F., Gaïffas, S., and Guilloux, A. (2011). Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(4):1171–1196.

Dahmen, W., DeVore, R., and Scherer, K. (1980). Multi-dimensional spline approximation. *SIAM Journal on Numerical Analysis*, 17(3):pp. 380–402.

Duane, J. T. (1964). Learning curve approach to reliability monitoring. *IEEE Transactions on Aerospace*, 2(2):563–566.

Goel, A. L. and Okumoto, K. (1979). Time-dependent error-detection rate model for software reliability and other performance measures. *Reliability, IEEE Transactions on*, R-28(3):206–211.

Juditsky, A., Lepski, O., and Tsybakov, A. (2009). Nonparametric estimation of composite functions. *The Annals of Statistics*, 37(3):1360–1404.

Krishnamurthy, K., Raginsky, M., and Willett, R. (2010). Multiscale photon-limited spectral image reconstruction. *SIAM Journal on Imaging Sciences*, 3:619–645.

Massart, P. (2003). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin/Heidelberg. École d'été de Probabilités de Saint-Flour.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126:103–153.

Yamada, S., Mitsuru, O., and Osaki, S. (1983). S-shaped reliability growth modeling for software error detection. *IEEE Transactions on Reliability*, 32(5):475–484.

CHAPITRE 3

Estimation of the transition density of a Markov chain

This chapter is a slightly modified version of the paper *Estimation of the transition density of a Markov chain* to appear in *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques.*

Abstract

We present two data-driven procedures to estimate the transition density of an homogeneous Markov chain. The first yields a piecewise constant estimator on a suitable random partition. By using an Hellinger-type loss, we establish non-asymptotic risk bounds for our estimator when the square root of the transition density belongs to possibly inhomogeneous Besov spaces with possibly small regularity index. Some simulations are also provided. The second procedure is of theoretical interest and leads to a general model selection theorem from which we derive rates of convergence over a very wide range of possibly inhomogeneous and anisotropic Besov spaces. We also investigate the rates that can be achieved under structural assumptions on the transition density.

1. Introduction

Consider a time-homogeneous Markov chain $(X_i)_{i\in\mathbb{N}}$ defined on an abstract probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with values in the measured space $(\mathbb{X}, \mathcal{F}, \mu)$. We assume that for each $x \in \mathbb{X}$, the conditional law $\mathcal{L}(X_{i+1} \mid X_i = x)$ admits a density $s(x, \cdot)$ with respect to μ. Our aim is to estimate the transition density $(x, y) \mapsto s(x, y)$ on a subset $A = A_1 \times A_2$ of $\mathbb{X}^2$ from the observations $X_0, \ldots, X_n$.

Many papers are devoted to this statistical setting. A popular method to build an estimator of s is to divide an estimator of the joint density of (X_i, X_{i+1}) by an estimator of the density of X_i. The resulting estimator is called a quotient estimator. Roussas (1969), Athreya and Atuncar (1998) considered Kernel estimators for the densities of X_i and (X_i, X_{i+1}). They proved consistence and asymptotic normality of the quotient estimator. Other properties of this estimator were established: Roussas (1991), Dorea (2002) showed strong consistency, Basu and Sahoo (1998) proved a Berry-Essen type theorem and Doukhan and Ghindès (1983) bounded from above the integrated quadratic risk under Sobolev constraints. Clémencon (2000) investigated

the minimax rates when $A = [0,1]^2$, $\mathbb{X}^2 = \mathbb{R}^2$. Given two smoothness classes $\mathscr{F}_1$ and $\mathscr{F}_2$ of real valued functions on $[0,1]^2$ and $[0,1]$ respectively (balls of Besov spaces), he established the lower bounds over the class

$$\mathscr{F} = \left\{ \varphi,\ \forall x, y \in [0,1], \varphi(x,y) = \frac{\varphi_1(x,y)}{\varphi_2(x)},\ (\varphi_1, \varphi_2) \in \mathscr{F}_1 \times \mathscr{F}_2 \right\}.$$

He developed a method based on wavelet thresholding to estimate the densities of X_i and (X_i, X_{i+1}) and showed that the quotient estimator of s is quasi-optimal in the sense that the minimax rates are achieved up to possible logarithmic factors. Lacour (2008, 2012) used model selection via penalization to construct estimates of the densities. The resulting quotient estimator reaches the minimax rates over $\mathscr{F}$ when $\mathscr{F}_1$ and $\mathscr{F}_2$ are balls of homogeneous (but possibly anisotropic) Besov spaces on $[0,1]^2$ and $[0,1]$ respectively.

The previous rates of convergence depend on the smoothness properties of the densities of X_i and (X_i, X_{i+1}). In the favourable case where $X_0, \ldots, X_n$ are drawn from a stationary Markov chain (with stationary density f), the rates depend on the smoothness properties of f or more precisely on the restriction of f to A_1. This function may however be less regular than the target function s. We refer for instance to Section 5.4.1 of Clémencon (2000) for an example of a Doeblin recurrent Markov chain where the stationary density f is discontinuous on $[0,1]$ although s is constant on $[0,1]^2$. Therefore, these estimators may converge slowly even if s is smooth, which is problematic.

This issue was overcome in several papers. Clémencon (2000) proposed a second procedure, based on wavelets and an analogy with the regression setting. He computed the lower bounds of minimax rates when the restriction of s on $[0,1]^2$ belongs to balls of some (possibly inhomogenous) Besov spaces and proved that its estimator achieves these rates up to a possible logarithmic factor. Lacour (2007) established lower bound over balls of some (homogenous but possibly anisotropic) Besov spaces. By minimizing a penalized contrast inspired from the least-squares, she obtained a model selection theorem from which she deduced that her estimator reaches the minimax rates when $A = [0,1]^2$, $\mathbb{X}^2 = \mathbb{R}^2$. With a similar procedure, Akakpo and Lacour (2011) obtained the usual rates of convergence over balls of possibly anisotropic and inhomogeneous Besov spaces (when $\mathbb{X}^2 = A = [0,1]^{2d}$). Very recently, Birgé (2012) proposed a procedure based on robust testing to establish a general oracle inequality. The expected rates of convergence can be deduced from this inequality when $\sqrt{s}$ belongs to balls of possibly anisotropic and inhomogeneous Besov spaces.

These authors have used different losses in order to evaluate the performance of their estimators. In each of these papers, the risk of an estimator $\hat{s}$ is of the form $\mathbb{E}\left[\delta^2(s\mathbb{1}_A, \hat{s})\right]$ where $\mathbb{1}_A$ denotes the indicator function of the subset A and δ a suitable distance. Lacour (2007), Akakpo and Lacour (2011) considered the space $\mathbb{L}^2(\mathbb{X}^2, M)$ of square integrable functions on $\mathbb{X}^2$ equipped with the random product measure $M = \lambda_n \otimes \mu$ where $\lambda_n = n^{-1} \sum_{i=0}^{n-1} \delta_{X_i}$ and used the distance defined for $f, f' \in \mathbb{L}^2(\mathbb{X}^2, M)$ by

$$\delta^2(f, f') = \frac{1}{n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \left(f(X_i, y) - f'(X_i, y)\right)^2 \mathrm{d}\mu(y).$$

Birgé (2012) considered the cone $\mathbb{L}^1_+(\mathbb{X}^2, \mu \otimes \mu)$ of non-negative integrable functions and used

the deterministic Hellinger-type distance defined for $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, \mu \otimes \mu)$ by

$$\delta^2(f, f') = \frac{1}{2} \int_{\mathbb{X}^2} \left(\sqrt{f(x,y)} - \sqrt{f'(x,y)} \right)^2 \mathrm{d}\mu(x)\, \mathrm{d}\mu(y).$$

These approaches, which often rely on the loss that is used, require the knowledge (or at least a suitable estimation) of various quantities depending on the unknown s, such as the supremum norm of s, or on a positive lower bound, either on the stationary density, or on $k^{-1} \sum_{j=1}^{k} s^{(l+j)}$ for some $k \geq 1$, $l \geq 0$ where $s^{(l+j)}(x, \cdot)$ is the density of the conditional law $\mathcal{L}(X_{l+j} \mid X_0 = x)$. Unfortunately, these quantities not only influence the way the estimators are built but also their performances since they are involved in the risk bounds. In this work, we shall rather consider the distance H (corresponding to an analogue of the random $\mathbb{L}^2$ loss above) defined on the cone $\mathbb{L}_+^1(\mathbb{X}^2, M)$ of integrable and non-negative functions by

$$H^2(f, f') = \frac{1}{2n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \left(\sqrt{f(X_i, y)} - \sqrt{f'(X_i, y)} \right)^2 \mathrm{d}\mu(y) \quad \text{for all } f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M).$$

For such a loss, we shall show that our estimators satisfy an oracle-type inequality under very weak assumptions on the Markov chain. A connection with the usual deterministic Hellinger-type loss will be done under a posteriori assumptions on the chain, and hence, independently of the construction of the estimator.

Our estimation strategy can be viewed as a mix between an approach based on the minimization of a contrast and an approach based on robust tests. Estimation procedures based on tests started in the seventies with Lucien Lecam and Lucien Birgé (Le Cam (1973, 1975); Birgé (1983, 1984a,b)). More recently, Birgé (2006) presented a powerful device to establish general oracle inequalities from robust tests. It was used in our statistical setting in Birgé (2012), in many others in Birgé (2007, 2013) and in Chapter 2 of the present thesis. We make two contributions to this area. Firstly, we provide a new test for our statistical setting. This test is based on a variational formula inspired from Baraud (2011) and differs from the one of Birgé (2012). Secondly, we shall study procedures that are quite far from the original one of Birgé (2006). Let us explain why.

The procedure of Birgé (2006) depends on a suitable net, the construction of which is usually abstract, making thus the estimator impossible to build in practice. In the favourable cases where the net can be made explicit, the procedure is anyway too complex to be implemented (see for instance Section 3.4.2 of Birgé (2007)). This procedure was afterwards adapted to estimators selection in Baraud and Birgé (2009) (for histogram type estimators) and in Baraud (2011) (for more general estimators). The complexity of their algorithms is of order the square of the cardinality of the family and are thus implementable when this family is not too large. In particular, given a family of histogram type estimators $\{\hat{s}_m, m \in \mathcal{M}\}$, these two procedures are interesting in practice when $\mathcal{M}$ is a collection of regular partitions (namely when all its elements have same Lebesgue measure) but become unfortunately numerically intractable for richer collections. In this work, we tackle this issue by proposing a new way of selecting among a family of piecewise constant estimators when the collection $\mathcal{M}$ ensues from the adaptive approximation algorithm of DeVore and Yu (1990).

We present this procedure in the first part of the chapter. It yields a piecewise constant estimator on a data-driven partition that satisfies an oracle-type inequality from which we shall

deduce uniform rates of convergence over balls of (possibly) inhomogeneous Besov spaces with small regularity indices. These rates coincide, up to a possible logarithmic factor to the usual ones over such classes. Finally, we carry out numerical simulations to compare our estimator with the one of Akakpo and Lacour (2011).

In the second part of this chapter, we are interested in obtaining stronger theoretical results for our statistical problem. We put aside the practical considerations to focus on the construction of an estimator that satisfies a general model selection theorem. Such an estimator should be considered as a benchmark for what theoretically feasible. We deduce rates of convergence over a large range of anisotropic and inhomogeneous Besov spaces on $[0,1]^{2d}$. We shall also consider other kinds of assumptions on the transition density. We shall assume that s belongs to classes of functions satisfying structural assumptions and for which faster rates of convergence can be achieved. This approach was developed by Juditsky et al. (2009) (in the Gaussian white noise model) and by Baraud and Birgé (2011) (in more statistical settings) to avoid the curse of dimensionality. More precisely, Baraud and Birgé (2011) showed that these rates can be deduced from a general model selection theorem, which strengthen its theoretical interest. This strategy was used in Chapter 2 to establish risk bounds over many classes of functions for Poisson processes with covariates. In this chapter, we shall use these assumptions to obtain faster rates of convergence for autoregressive Markov chains (whose conditional variance may not be constant).

This chapter is organized as follows. The first procedure, which selects among piecewise constant estimators is presented and theoretically studied in Section 2. In Section 3, we carry out a simulation study and compare our estimator with the one of Akakpo and Lacour (2011). The practical implementation of this procedure is quite technical and will therefore be delayed in the appendix, in Section 5. In Section 4, we establish theoretical results by using our second procedure. The proofs are postponed to Section 6.

Let us introduce some notations that will be used all along the chapter. The number $x \vee y$ (respectively $x \wedge y$) stands for $\max(x,y)$ (respectively $\min(x,y)$) and x_+ stands for $x \vee 0$. We set $\mathbb{N}^\star = \mathbb{N} \setminus \{0\}$. For (E,d) a metric space, $x \in E$ and $A \subset E$, the distance between x and A is denoted by $d(x,A) = \inf_{a \in A} d(x,a)$. The indicator function of a subset A is denoted by $\mathbb{1}_A$ and the restriction of a function f to A by $f_{|A}$. For all real valued function f on E, $\|f\|_\infty$ stands for $\sup_{x \in E} |f(x)|$. The cardinality of a finite set A is denoted by $|A|$. The notations C,C',C''... are for the constants. The constants C,C',C''... may change from line to line.

2. Selecting among piecewise constant estimators

Throughout this section, we assume that $\mathbb{X} = \mathbb{R}^d$, $A = [0,1]^{2d}$, $\mu([0,1]^d) = 1$ and $n > 3$.

2.1. Preliminary estimators. Given a (finite) partition m of $[0,1]^{2d}$, a simple way to estimate s on $[0,1]^{2d}$ is to consider the piecewise constant estimator on the elements of m defined by

$$\hat{s}_m = \sum_{K \in m} \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x)\, \mathrm{d}\mu(x)} \mathbb{1}_K. \tag{1}$$

In the above definition, the denominator $\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x)\, \mathrm{d}\mu(x)$ may be equal to 0 for some sets K, in which case the numerator $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) = 0$ as well, and we shall use the convention $0/0 = 0$.

We now bound from above the risk of this estimator. We set

$$V_m = \left\{ \sum_{K \in m} a_K \mathbb{1}_K,\, \forall K \in m,\, a_K \in [0, +\infty) \right\}$$

and prove the following.

Proposition 1. *For all finite partition m of $[0,1]^{2d}$,*

$$C\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s}_m)\right] \leq \mathbb{E}\left[H^2(s\mathbb{1}_A, V_m)\right] + \frac{1+\log n}{n}|m|$$

where $C = 1/(4 + \log 2)$.

Up to a constant, the risk of $\hat{s}_m$ is bounded by a sum of two terms. The first one corresponds to the approximation term whereas the second one corresponds to the estimation term.

An analogue upper bound on the empirical quadratic risk of this estimator may be found in Chapter 4 of Akakpo (2009). Her bound requires several assumptions on the partition m and the Markov chain although the present one requires none. However, unlike hers, we lose a logarithmic term.

2.2. Definition of the partitions. In this section, we shall deal with special choice of partitions m. More precisely, we consider the family of partitions defined by using the recursive algorithm developed in DeVore and Yu (1990). For $j \in \mathbb{N}$, we consider the set

$$\mathcal{L}_j = \left\{ \mathbf{l} = (l_1, \ldots, l_{2d}) \in \mathbb{N}^{2d},\, 1 \leq l_i \leq 2^j \text{ for } 1 \leq i \leq 2d \right\}$$

and define for all $\mathbf{l} = (l_1, \ldots, l_{2d}) \in \mathcal{L}_j$,

$$\forall i \in \{1, \ldots, 2d\}, \quad I_j(l_i) = \begin{cases} \left[\frac{l_i - 1}{2^j}, \frac{l_i}{2^j}\right) & \text{if } l_i < 2^j \\ \left[\frac{l_i - 1}{2^j}, 1\right] & \text{if } l_i = 2^j. \end{cases}$$

We then introduce the cube $K_{j,\mathbf{l}} = \prod_{i=1}^{2d} I_j(l_i)$ and set $\mathcal{K}_j = \{K_{j,\mathbf{l}},\, \mathbf{l} \in \mathcal{L}_j\}$.

The algorithm starts with $[0,1]^{2d}$. At each step, it gets a partition of $[0,1]^{2d}$ into a finite family of disjoint cubes of the form $K_{j,\mathbf{l}}$. For any such cube, one decides to divide it into the 4^d elements of $\mathcal{K}_{j+1}$ which are contained in it, or not. The set of all such partitions that can be constructed in less than ℓ steps is denoted by $\mathcal{M}_\ell$. We set $\mathcal{M}_\infty = \cup_{\ell \geq 1} \mathcal{M}_\ell$. Two examples of partitions are illustrated in Figure 3.1 (for $d = 1$).

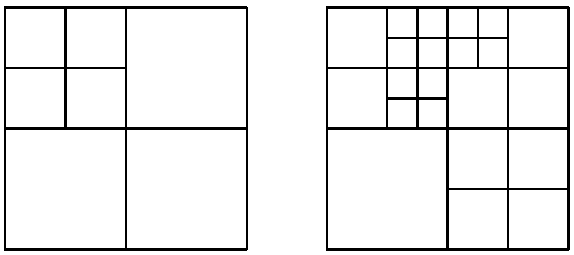


Figure 3.1: Left: example of a partition of $\mathcal{M}_2$. Right: example of a partition of $\mathcal{M}_3$.

2.3. The selection rule. Given $\ell \in \mathbb{N}^\star \cup \{\infty\}$, the aim of this section is to select an estimator among the family $\{\hat{s}_m,\, m \in \mathcal{M}_\ell\}$.

For any $K \in \cup_{m\in\mathcal{M}_\ell} m$ and any partition $m' \in \mathcal{M}_\ell$, let $m' \vee K$ be the partition of K defined by

$$m' \vee K = \{K' \cap K,\, K' \in m',\, K \cap K' \neq \emptyset\}.$$

Let L be a positive number and pen be the non-negative map defined by

$$\mathrm{pen}\left(m' \vee K\right) = L\frac{|m' \vee K|\log n}{n} \quad \text{for all } m' \in \mathcal{M}_\ell \text{ and } K \in \cup_{m\in\mathcal{M}_\ell} m.$$

This definition implies in particular that

$$\mathrm{pen}(m) = L\frac{|m|\log n}{n} \quad \text{for all partition } m \in \mathcal{M}_\ell.$$

Let us set $\alpha = (1 - 1/\sqrt{2})/2$ and for all $f, f' \in \mathbb{L}^1_+(\mathbb{X}^2, M)$,

$$\begin{aligned} T(f,f') \;=\; & \frac{1}{2n\sqrt{2}}\sum_{i=0}^{n-1}\int_{\mathbb{X}} \sqrt{f(X_i,y)+f'(X_i,y)}\left(\sqrt{f'(X_i,y)}-\sqrt{f(X_i,y)}\right)\mathrm{d}\mu(y) \\ & + \frac{1}{n\sqrt{2}}\sum_{i=0}^{n-1}\frac{\sqrt{f'(X_i,X_{i+1})}-\sqrt{f(X_i,X_{i+1})}}{\sqrt{f(X_i,X_{i+1})+f'(X_i,X_{i+1})}} \\ & + \frac{1}{2n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\left(f(X_i,y)-f'(X_i,y)\right)\mathrm{d}\mu(y). \end{aligned}$$

We define γ for $m \in \mathcal{M}_\ell$ by

$$\gamma(m) = \left\{\sum_{K\in m}\sup_{m'\in\mathcal{M}_\ell}\left[\alpha H^2\left(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'}\mathbb{1}_K\right) + T\left(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'}\mathbb{1}_K\right) - \mathrm{pen}(m'\vee K)\right]\right\} + 2\mathrm{pen}(m).$$

Finally, we select $\hat{m}$ among $\mathcal{M}_\ell$ as any partition satisfying

$$\gamma(\hat{m}) \leq \inf_{m\in\mathcal{M}_\ell}\gamma(m) + \frac{1}{n} \tag{2}$$

and consider the resulting estimator $\hat{s} = \hat{s}_{\hat{m}}$.

Remarks. The estimator $\hat{s} = \hat{s}(L, \ell)$ depends on the choices of two quantities $L > 0$, $\ell \in \mathbb{N}^\star \cup \{\infty\}$. We shall see in the next section that L can be chosen as an universal numerical constant. As to ℓ, from a theoretical point of view, it can be chosen as $\ell = \infty$. In practice, we recommend to take it as large as possible. Nevertheless, the larger ℓ, the longer it takes to compute the estimator. A practical algorithm in view of computing $\hat{m}$ will be detailed in the appendix.

The selection procedure we use may look somewhat unusual. It can be seen as a mix between a procedure based on a contrast function (which is usually easy to implement) and a procedure based on a robust test (the functional $T(f, f') = -T(f', f)$, which can be seen as a robust test between f, f', will allow us to obtain risk bounds with respect to a Hellinger-type distance). This functional is inspired from the variational formula for the Hellinger affinity described in Section 2 of Baraud (2011).

In the literature, procedures based on a robust test are usually based on the minimization of a functional D known as plausibility index. In our context, D would be

$$D(m) = \sup \left\{ H^2(\hat{s}_m, \hat{s}_{m'}),\ m' \in \mathcal{M}_\ell, T(\hat{s}_m, \hat{s}_{m'}) \geq \mathrm{pen}(m') - \mathrm{pen}(m) \right\}$$

and the estimator would be defined by $\hat{s}_{\tilde{m}}$ where $\tilde{m}$ minimizes $D(m)$ over $m \in \mathcal{M}_\ell$. The computation of $D(m)$ is unfortunately numerically intractable, which implies that $\tilde{m}$ is purely theoretical. The computation of the supremum is a constraint optimization problem and the Lagrange multipliers suggest to replace $D(m)$ by

$$\gamma_1(m) = \sup_{m' \in \mathcal{M}_\ell} \left[\alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \mathrm{pen}(m') \right] + \mathrm{pen}(m).$$

The contrast γ can be interpreted as being a modification of γ_1 whose minimum can be found in practice. The minimums of γ, γ_1 and D may not be equal, but it can be shown that they possess similar statistical properties.

2.4. An oracle inequality. The main result of this section is the following.

Theorem 2. *There exists an universal constant $L_0 > 0$ such that, for all $L \geq L_0$, $\ell \in \mathbb{N}^\star \cup \{\infty\}$, the estimator $\hat{s} = \hat{s}(L, \ell)$ satisfies*

$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s} \right) \right] \leq \inf_{m \in \mathcal{M}_\ell} \left\{ \mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m \right) \right] + \mathrm{pen}(m) \right\} \tag{3}$$

where C is an universal positive constant.

In the literature, oracle inequalities with a random quadratic loss for piecewise constant estimators have been obtained in Lacour (2007) and Akakpo and Lacour (2011). Their procedures require a priori assumptions on the transition density and the Markov chain although ours requires none (except homogeneity). However, unlike theirs, our risk bound involves an extra logarithmic term. We do not know whether this term is necessary or not.

In the proof, we obtain an upper bound for L_0 which is unfortunately very rough and useless in practice. It seems difficult to obtain a sharp bound on L_0 from the theory and we have rather carried out a simulation study in order to tune L_0 (see Section 3).

2.5. Risk bounds with respect to a deterministic loss. Although the distance H is natural, we are interested in controlling the risk associated to a deterministic distance. To do so, we shall make a posteriori assumptions on the Markov chain.

Assumption 1. *The sequence $(X_i)_{i\geq 0}$ is stationary and admits a stationary density φ with respect to the Lebesgue measure μ on $\mathbb{R}^d$. There exists $\kappa_0 > 0$ such that $\varphi(x) \geq \kappa_0$ for all $x \in [0,1]^d$.*

We introduce $\mathbb{L}_+^1\left([0,1]^{2d}, (\varphi\cdot\mu)\otimes\mu\right)$ the cone of integrable and non-negative functions on $[0,1]^{2d}$ with respect to the product measure $(\varphi\cdot\mu)\otimes\mu$. We endow $\mathbb{L}_+^1([0,1]^{2d}, (\varphi\cdot\mu)\otimes\mu)$ with the distance h defined by

$$\forall f, f' \in \mathbb{L}_+^1\left([0,1]^{2d}, (\varphi\cdot\mu)\otimes\mu\right), \quad h^2(f,f') = \frac{1}{2}\int_{[0,1]^{2d}} \left(\sqrt{f(x,y)} - \sqrt{f'(x,y)}\right)^2 \varphi(x)\,\mathrm{d}x\,\mathrm{d}y.$$

In our results, we shall need the β-mixing properties of the Markov chain. We set for all $q \in \mathbb{N}^\star$

$$\beta_q = \int_{\mathbb{R}^{2d}} |s^{(q)}(x,y) - \varphi(y)|\varphi(x)\,\mathrm{d}x\,\mathrm{d}y$$

where $s^{(q)}(x,\cdot)$ is the density of the conditional law $\mathcal{L}(X_q \mid X_0 = x)$ with respect to the Lebesgue measure. We refer to Doukhan (1994) and Bradley (2005) for more details on the β-mixing coefficients.

Theorem 3. *Under Assumption 1, the estimator $\hat{s}$ built in Section 2.3 with $\ell \in \mathbb{N}^\star$ and $L \geq L_0$, satisfies*

$$C\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \inf_{m\in\mathcal{M}_\ell}\left\{h^2\left(s\mathbb{1}_A, V_m\right) + \mathrm{pen}(m)\right\} + \frac{R_n(\ell)}{n}$$

where

$$R_n(\ell) = n2^{3\ell d}\inf_{1\leq q\leq n}\left\{\exp\left(-\frac{\kappa_0}{10}\frac{n}{q2^{\ell d}}\right) + n\beta_q/q\right\} \tag{4}$$

and where C is an universal positive constant.

This result is interesting when the remainder term $R_n(\ell)/n$ is small enough, that is when $2^{\ell d}$ is small compared to n and when the sequence $(\beta_q)_{q\geq 1}$ goes to 0 fast enough. More precisely, $R_n(\ell)$ can be bounded independently of n, ℓ whenever ℓ, d, n and the β_q coefficients satisfy the following.

- If the chain is geometrically β-mixing, that is if there exists $b_1 > 0$ such that $\beta_q \leq e^{-b_1 q}$, then

$$R_n(\ell) \leq n^2 2^{3\ell d+1}\left[\exp(-b_1 n) + \exp\left(-\frac{\kappa_0}{10}\frac{n}{2^{\ell d}}\right) + \exp\left(-\sqrt{\frac{\kappa_0 b_1}{40}\frac{n}{2^{\ell d}}}\right)\right].$$

  In particular, if ℓ, d, n are such that $2^{\ell d} \leq n/\log^3 n$, $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_1.

- If the chain is arithmetically β-mixing, that is if there exists $b_2 > 0$ such that $\beta_q \leq q^{-b_2}$, then
$$R_n(\ell) \leq \frac{C'(b_2)}{\kappa_0^{b_2+1}} \frac{2^{(4+b_2)\ell d} \log^{b_2+1}\left(1 + \frac{\kappa_0 n}{2^{\ell d}}\right)}{n^{b_2-1}}$$
where $C'(b_2)$ depends only on b_2. Consequently, if $2^{\ell d} \leq n^{1-\zeta}/\log n$ and $b_2 \geq 5/\zeta - 4$ for $\zeta \in (0,1)$, $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_2.

2.6. Rates of convergence. The aim of this section is to obtain uniform risk bounds over classes of smooth transition densities for our estimator.

2.6.1. Hölder spaces. Given $\sigma \in (0,1]$, we say that a function f belongs to the Hölder space $\mathcal{H}^\sigma([0,1]^{2d})$ if there exists $|f|_\sigma \in \mathbb{R}_+$ such that for all $(x_1,\dots,x_{2d}) \in [0,1]^{2d}$ and all $1 \leq j \leq 2d$, the functions $f_j(\cdot) = f(x_1,\dots,x_{j-1},\cdot,x_{j+1},\dots,x_{2d})$ satisfy
$$|f_j(x) - f_j(y)| \leq |f|_\sigma |x-y|^\sigma \quad \text{for all } x,y \in [0,1].$$
When the restriction of $\sqrt{s}$ to $A = [0,1]^{2d}$ is Hölderian, we deduce from (3) the following.

Corollary 1. *For all $\sigma \in (0,1]$ and $\sqrt{s}|_A \in \mathcal{H}^\sigma([0,1]^{2d})$, the estimator $\hat{s} = \hat{s}(L_0,\infty)$ satisfies*
$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \left(d\left|\sqrt{s}|_A\right|_\sigma\right)^{\frac{2d}{d+\sigma}} \left(\frac{\log n}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n}$$
where C is an universal positive constant.

2.6.2. Besov spaces. A thinner way to measure the smoothness of the transition density is to assume that $\sqrt{s}|_A$ belongs to a Besov space. We refer to Section 3 of DeVore and Yu (1990) for a definition of this space. We say that the Besov space $\mathscr{B}_q^\sigma(\mathbb{L}^p([0,1]^{2d}))$ is homogeneous when $p \geq 2$ and inhomogeneous otherwise. We set for all $p \in (1,+\infty)$ and $\sigma \in (0,1)$,
$$\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d})) = \begin{cases} \mathscr{B}_p^\sigma(\mathbb{L}^p([0,1]^{2d})) & \text{if } p \in (1,2) \\ \mathscr{B}_\infty^\sigma(\mathbb{L}^p([0,1]^{2d})) & \text{if } p \in [2,+\infty), \end{cases}$$
and denote by $|\cdot|_{p,\sigma}$ the semi norm of $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$. We make the following assumption to deduce from (3) risk bounds over these spaces.

Assumption 2. *There exists $\kappa > 0$ such that for all $i \in \{0,\dots,n-1\}$, X_i admits a density φ_i with respect to the Lebesgue measure μ such that $\varphi_i(x) \leq \kappa$ for all $x \in [0,1]^d$.*

Note that we do not require that the chain be either stationary or mixing.

Let $\left(\mathbb{L}^2([0,1]^{2d}, \mu\otimes\mu), d_2\right)$, be the metric space of square integrable functions on $[0,1]^{2d}$ with respect to the Lebesgue measure. Under the above assumption, we deduce from (3) that
$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] \leq \inf_{m\in\mathcal{M}_\ell} \left\{ \kappa d_2^2\left(\sqrt{s}|_A, V_m\right) + L_0 \frac{|m|\log n}{n} \right\}.$$
When $\sqrt{s}|_A$ belongs to a Besov space, the right-hand side of this inequality can be upper bounded thanks to the approximation theorems of DeVore and Yu (1990).

Corollary 2. *Suppose that Assumption 2 holds. For all* $p \in (2d/(d+1), +\infty)$, $\sigma \in (2d(1/p - 1/2)_+, 1)$ *and* $\sqrt{s}|_A \in \mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$, *the estimator* $\hat{s} = \hat{s}(L_0, \infty)$ *satisfies*

$$C'\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}} \left(\frac{\log n}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} \tag{5}$$

where $C' > 0$ *depends only on* κ,σ,d,p.

More precisely, it is shown in the proof that the estimators $\hat{s} = \hat{s}(L_0, \ell)$ satisfy (5) when ℓ is large enough (when $\ell \geq d^{-1}(\log 2)^{-1} \log n$).

Rates of convergence for the deterministic loss h can be established by using Theorem 3 instead of Theorem 2. For instance, if the chain is geometrically β-mixing, we may choose ℓ the smallest integer larger than $d^{-1}(\log 2)^{-1} \log(n/\log^3 n)$, in which case the estimator $\hat{s} = \hat{s}(L_0, \ell)$ achieves the rate $(\log n/n)^{\sigma/(\sigma+d)}$ over the Besov spaces $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$, $p \in (2d/(d+1), +\infty)$, $\sigma \in (\sigma_1(p,d), 1)$ where

$$\sigma_1(p,d) = \frac{d}{4}\left(-1 + 4\left(1/p - 1/2\right)_+ + \sqrt{1 + 24\left(1/p - 1/2\right)_+ + 16\left(1/p - 1/2\right)_+^2}\right).$$

If the chain is arithmetically β-mixing with $b_q \leq q^{-6}$, choosing ℓ the smallest integer larger than $d^{-1}(2\log 2)^{-1} \log(n/\log n)$ allows us to recover the same rate of convergence when $\sigma \in (\sigma_2(p,d), 1)$ where

$$\sigma_2(p,d) = d\left(\left(1/p - 1/2\right)_+ + \sqrt{2(1/p - 1/2)_+ + (1/p - 1/2)_+^2}\right).$$

We refer the reader to Section 6.6 for a proof of these two results.

In the literature, Lacour (2007) obtained a rate of order $n^{-\sigma/(\sigma+1)}$ over $\mathscr{B}^\sigma(\mathbb{L}^2([0,1]^2))$, which is slightly faster but her approach prevents her to deal with inhomogeneous Besov spaces and requires the prior knowledge of a suitable upper bound on the supremum norm of s. As far as we know, the rates that have been established in the other papers hold only when $\sigma > 1$.

3. Simulations

In this section, we present a simulation study to evaluate the performance of our estimator in practice. We shall simulate several Markov chains and estimate their transition densities by using our procedure.

The program we have used to compute the estimator is available at

http://math.unice.fr/~msart/

3.1. Examples of Markov chains.

We consider Markov chains of the form

$$X_{k+1} = F(X_k, U_k)$$

where F is some known function and where U_k is a random variable independent of $(X_0, \ldots, X_k)$.

For the sake of comparison, we begin to deal with examples that have already been considered in the simulation study of Akakpo and Lacour (2011). In each of these examples, U_k is a standard Gaussian random variable.

Example 1. $X_{k+1} = 0.5X_k + (1 + U_k)/4$

Example 2. $X_{k+1} = 12^{-1}\left(6 + \sin(12X_k - 6) + (\cos(X_k - 6) + 3)U_k\right)$

Example 3.

$$X_{k+1} = \frac{1}{3}\left(X_k + 1\right) + \left(\frac{1}{9} - \frac{1}{23}\left(\frac{1}{2}\beta(5X_i/3, 4, 4) + \frac{1}{20}\beta\left((5X_i - 2)/3, 400, 400\right)\right)\right)U_k$$

where $\beta(\cdot, a, b)$ is the density of the β distribution with parameters a and b.

Example 4.

$$X_{k+1} = \frac{1}{4}\left(g(X_k) + 1\right) + \frac{1}{8}U_k$$

where g is defined by

$$g(x) = \frac{9\sqrt{2}}{4\sqrt{\pi}}\exp\left(-18(x - 1/2)^2\right) + \frac{9\sqrt{2}}{4\sqrt{\pi}}\exp\left(-162(x - 3/4)^2\right) \quad \textit{for all } x \in \mathbb{R}.$$

At first sight, Examples 1 and 2 may seem to be different than those of Akakpo and Lacour (2011). Actually, we just have rescaled the data in order to estimate on $[0, 1]^2$. The statistical problem is the same. According to Akakpo and Lacour (2011), we set p large ($p = 10^4$) and we estimate the transition densities of Examples 1, 2, 3 and 4 from $(X_p, \ldots, X_{n+p})$ so that the chain is approximatively stationary.

We also propose to consider the following examples. In Example 5, U_k is a centred Gaussian random variable with variance $1/2$, in Example 6, U_k admits the density

$$f(x) = \frac{5\sqrt{2}}{2\sqrt{\pi}}\left[\exp\left(-50(x - 1)^2\right) + \exp\left(-50x^2\right)\right]$$

with respect to the Lebesgue measure, and in Example 7, U_k is an exponential random variable with parameter 1.

Example 5. $X_{k+1} = 0.5X_k + (1 + U_k)/4$.

Example 6. $X_{k+1} = 0.5\left(X_k + U_k\right)$.

Example 7. $X_{k+1} = X_k/(50X_k + 1) + X_kU_k$.

We set $X_0 = 1/2$ and estimate s from $(X_0, \ldots, X_n)$. These last three Markov chains are not stationary. Their transition densities are rather isotropic and inhomogeneous. The transition density of Example 7 is unbounded.

In what follows, our selection rule will always be applied with $L = 0.03$ (whatever, ℓ, n and the Markov chain).

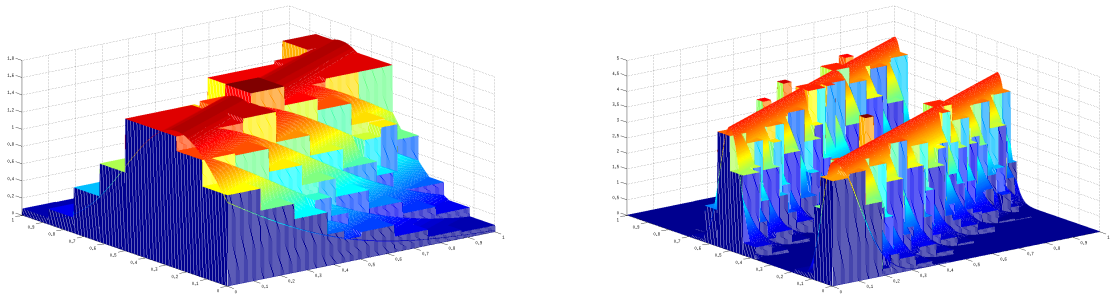
3.2. Choice of ℓ. We discuss the choice of ℓ by simulating the preceding examples with $n = 10^3$ and by applying our selection rule for each value of $\ell \in \{1, \dots, 10\}$. The results are summarized below.

ℓ	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
1	0.031	0.046	0.299	0.181	0.089	0.291	0.358
2	0.011	0.015	0.087	0.107	0.024	0.170	0.241
3	0.011	0.014	0.026	0.058	0.013	0.067	0.156
4	0.011	0.018	0.026	0.035	0.015	0.046	0.113
5	0.011	0.018	0.022	0.038	0.015	0.048	0.098
6	0.011	0.018	0.022	0.038	0.015	0.048	0.065
7	0.011	0.018	0.024	0.038	0.015	0.048	0.044
8	0.011	0.018	0.024	0.038	0.015	0.048	0.040
9	0.011	0.018	0.024	0.038	0.015	0.048	0.040
10	0.011	0.018	0.024	0.038	0.015	0.048	0.040

Figure 3.2: Hellinger risk $H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})$.

When ℓ grows up, the risk of our estimator tends to decrease and then stabilize. The best choice of ℓ is obviously unknown in practice but this array shows that a good way for choosing ℓ is to take it as large as possible. This is theoretically justified by Theorem 2 since the right-hand side of inequality (3) is a non-increasing function of ℓ.

3.3. An illustration. We apply our procedure for Examples 1 and 6 with $n = 10^4$, $\ell = 7$. We get two estimators and draw them with the corresponding transition density in Figure 3.3.



Example 1. Example 6.

Figure 3.3: Estimator and transition density.

This shows that the selected partition is thinner (respectively wider) to the points where the transition density is changing rapidly (respectively slowly), and is thus rather well adapted to the target function s.

3.4. Comparison with other procedures. In this section, we compare our selection rule with the oracle estimator and with the piecewise constant estimator of Akakpo and Lacour (2011).

The procedure of Akakpo and Lacour (2011) amounts to selecting an estimator among $\{\hat{s}_m, m \in \mathcal{M}'\}$ where $\hat{s}_m$ is defined by (1) and where $\mathcal{M}'$ is a collection of irregular partitions on $[0,1]^2$. Precisely, with their notations, we apply it with $J_\star = 5$, $\mathrm{pen}(m) = 3\|s_{\mathbf{I}A}\|_\infty |m|/n$ and with $\mathrm{pen}(m) = 3\|\hat{s}_{m^\bullet}\|_\infty |m|/n$ where $m^\bullet$ is a partition suitably chosen (following the recommendations of Akakpo and Lacour (2011), that is $J_\bullet = 3$). These two estimators are denoted by $\hat{s}^{(1)}$ and $\hat{s}^{(2)}$ respectively. Notice that these penalties, which are used in their simulation study, are not the ones prescribed by their theory. Their theoretical penalties also depend on a positive lower bound on the stationary density.

We denote by $\hat{s}^{(0)}$ the oracle estimator, that is the estimator defined as being a minimizer of the map $m \mapsto H^2(s1_{[0,1]^2}, \hat{s}_m)$ for $m \in \mathcal{M}_7$. This estimator is the best estimator of the family $\{\hat{s}_m,\, m \in \mathcal{M}_7\}$ and is known since the data are simulated. We consider the random variables

$$\mathcal{R}_i = \frac{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})}{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(i)})} \quad \text{for } i = 1, 2$$

and denote by $q_0(\alpha)$ the α-quantile of $\mathcal{R}_0$. Results obtained are given in Figure 4.1.

	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})]$	0.011	0.017	0.022	0.038	0.018	0.052	0.049
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(0)})]$	0.007	0.011	0.015	0.028	0.012	0.037	0.041
$q_0(0.5)$	1.473	1.513	1.443	1.369	1.422	1.420	1.200
$q_0(0.75)$	1.698	1.627	1.557	1.440	1.575	1.481	1.244
$q_0(0.9)$	1.921	1.834	1.683	1.509	1.749	1.543	1.290
$q_0(0.95)$	2.113	1.965	1.770	1.558	1.839	1.590	1.317
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(1)})]$	0.017	0.018	0.028	0.058	0.024	0.103	-
$\mathbb{P}\ (\mathcal{R}_1 \leq 1)$	0.964	0.740	0.908	1	0.984	1	-
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(2)})]$	0.013	0.018	0.028	0.062	0.023	0.096	0.133
$\mathbb{P}\ (\mathcal{R}_2 \leq 1)$	0.832	0.748	0.928	1	0.948	1	1

Figure 3.4: Risks for simulated data with $n = 1000$ averaged over 250 samples.

3.5. Comparison with a quadratic empirical risk. In Akakpo and Lacour (2011), the risks of the estimators are evaluated with a empirical quadratic norm and we can also compare the performances of our estimator to theirs by using this risk.

To do so, let us denote by $\|\cdot\|_n$ the empirical quadratic norm defined by

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} f^2(X_i, x)\,\mathrm{d}x \quad \text{for all } f \in \mathbb{L}^2(\mathbb{R}^2, M)$$

and set for $i \in \{1, 2\}$,

$$\mathcal{R}'_i = \frac{\|s\mathbb{1}_{[0,1]^2} - \hat{s}\|_n^2}{\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(i)}\|_n^2}.$$

The results obtained are presented in Figure 4.2. They are very similar to those of Figure 4.1.

	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6	Ex 7
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}\|_n^2]$	0.064	0.108	0.229	0.319	0.116	0.528	2.82
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(1)}\|_n^2]$	0.147	0.133	0.257	0.423	0.205	0.743	-
$\mathbb{P}(\mathcal{R}_1' \leq 1)$	0.980	0.820	0.788	0.984	0.992	1	-
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(2)}\|_n^2]$	0.091	0.129	0.262	0.418	0.159	0.739	6.08
$\mathbb{P}(\mathcal{R}_2' \leq 1)$	0.864	0.780	0.792	0.980	0.940	1	1

Figure 3.5: Risks for simulated data with $n = 1000$ averaged over 250 samples.

4. A general procedure

In Section 2, we used our selection rule to establish the oracle inequality (3), from which we deduced rates of convergence over Besov spaces $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$ with σ lower than 1. We now aim at obtaining rates for more general spaces of functions. This includes Besov spaces with regularity index larger than 1 and spaces corresponding to structural assumptions on s. We propose a second procedure to reach this goal.

The Markov chain takes its values into $\mathbb{X}$ and we estimate s on a subset A of the form $A = A_1 \times A_2$. We always assume that $n > 3$.

4.1. Procedure and preliminary result. Our second procedure is defined as follows. Let $\alpha = (1 - 1/\sqrt{2})/2$, $L > 0$, S be an at most countable set of $\mathbb{L}_+^1(\mathbb{X}^2, M)$ and $\Delta_S \geq 1$ be a map on S.

We define the application $\wp$ on S by

$$\wp(f) = \sup_{f' \in S} \left[\alpha H^2(f, f') + T(f, f') - L\frac{\Delta_S(f')}{n} \right] + L\frac{\Delta_S(f)}{n} \quad \text{for all } f \in S.$$

We select $\hat{s}$ among S as any element of S satisfying

$$\wp(\hat{s}) \leq \inf_{f \in S} \wp(f) + \frac{1}{n}.$$

We prove the following.

Proposition 4. *Suppose that $f(x) = 0$ for all $f \in S$ and $x \in \mathbb{X}^2 \setminus A$ and that $\sum_{f \in S} e^{-\Delta_S(f)} \leq 1$. There exists an universal constant $L_0 > 0$ such that if $L \geq L_0$, the estimator $\hat{s}$ satisfies*

$$C\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \mathbb{E}\left[\inf_{f \in S}\left\{H^2(s\mathbb{1}_A, f) + L\frac{\Delta_S(f)}{n}\right\}\right] \tag{6}$$

where C is an universal positive constant.

4.2. A general model selection theorem. We shall deduce from the above proposition a model selection theorem by choosing suitably S. To do so, we consider the following assumption.

Assumption 3. *For all $i \in \{0, \dots, n-1\}$, X_i admits a density φ_i with respect to some known measure ν such that $\nu(A_1) = 1$. Moreover, there exists κ such that $\varphi_i(x) \leq \kappa$ for all $x \in A_1$ and $i \in \{0, \dots, n-1\}$.*

We define $\mathbb{L}^2(A, \nu \otimes \mu)$ the space of square integrable functions on A with respect to the product measure $\nu \otimes \mu$, and we endow it with its natural distance

$$d^2(f, f') = \int_A \big(f(x,y) - f'(x,y)\big)^2 \, \mathrm{d}\nu(x) \, \mathrm{d}\mu(y) \quad \text{for all } f, f' \in \mathbb{L}^2(A, \nu \otimes \mu).$$

Hereafter, a model V is a (non-trivial) finite dimensional linear space of $\mathbb{L}^2(A, \nu \otimes \mu)$.

Let us explain how to obtain a model selection theorem when Assumption 3 holds. Let $\mathbb{V}$ be a collection of models V and let $(\Delta(V))_{V \in \mathbb{V}}$ be a family of non-negative numbers such that $\sum_{V \in \mathbb{V}} e^{-\Delta(V)} \leq 1$. For each model $V \in \mathbb{V}$, we consider an orthonormal basis $(f_1, \dots, f_{\dim V})$ of V and set

$$T_V = \left\{ \sum_{i=1}^{\dim V} \alpha_i f_i, \, \alpha_i \in \frac{2}{\sqrt{n \dim V}} \mathbb{Z} \right\}.$$

We deduce from Lemma 5 of Birgé (2006) that the cardinality of

$$S_V = \left\{ f_+^2 \mathbb{1}_A, \, f \in T_V, \, d(f, 0) \leq 2 \right\}$$

is upper bounded by $|S_V| \leq (30n)^{\dim V/2}$. We then use the above procedure with $S = \cup_{V \in \mathbb{V}} S_V$ and

$$\Delta_S(f) = \inf_{\substack{V \in \mathbb{V} \\ S_V \ni f}} \left\{ \Delta(V) + (\dim V) \log(30n)/2 \right\} \quad \text{for all } f \in S.$$

This yields an estimator $\hat{s}$ such that

$$C' \mathbb{E}\left[H^2(s \mathbb{1}_A, \hat{s}) \right] \leq \inf_{V \in \mathbb{V}} \left\{ \kappa \left(\inf_{\substack{f \in T_V \\ d(f,0) \leq 2}} d^2\left(\sqrt{s}|_A, f \right) \right) + \frac{\Delta(V) + \dim(V) \log n}{n} \right\}$$

where C' is an universal positive constant. Since $d(\sqrt{s}|_A, 0) \leq 1$,

$$\inf_{\substack{f \in T_V \\ d(f,0) \leq 2}} d^2\left(\sqrt{s}|_A, f \right) = d^2\left(\sqrt{s}|_A, T_V \right).$$

For all $f' \in V$, there exists $f \in T_V$ such that $d^2(f, f') \leq n^{-1}$ and thus

$$d^2\left(\sqrt{s}|_A, T_V \right) \leq 2 d^2\left(\sqrt{s}|_A, V \right) + \frac{2}{n}.$$

Precisely, we have proved:

Theorem 5. *Suppose that Assumption 3 holds. Let $\mathbb{V}$ be an at most countable collection of models. Let $(\Delta(V))_{V\in\mathbb{V}}$ be a family of non-negative numbers such that*

$$\sum_{V\in\mathbb{V}} e^{-\Delta(V)} \leq 1.$$

There exists an estimator $\hat{s}$ such that

$$C\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ d^2\left(\sqrt{s}|_A, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\}$$

where $C > 0$ depends only on κ.

The condition $\sum_{\mathbb{V}\in\mathbb{V}} e^{-\Delta(V)} \leq 1$ can be interpreted as a (sub)probability on the collection $\mathbb{V}$. The more complex the family $\mathbb{V}$, the larger the weights $\Delta(V)$. When one can choose $\Delta(V)$ of order $\dim(V)$, which means that the family $\mathbb{V}$ of models does not contains too many models per dimension, the estimator $\hat{s}$ achieves the best trade-off (up to a constant) between the approximation and the variance terms.

This theorem holds under an assumption that is very mild and weaker than those of Lacour (2007), Akakpo and Lacour (2011) and Clémencon (2000). Birgé (2012) proved a general oracle inequality when there exist integers $k \geq 1$ and $l \geq 0$ and positive numbers ρ, ϱ such that

$$\varrho \leq \frac{1}{k}\sum_{j=1}^{k} s^{(l+j)}(x,y) \leq \rho \quad \text{for all } x, y \in \mathbb{X}$$

where the parameters k, l, ϱ are known. Our assumption is then satisfied for the Markov chain $(X_{l+1}, \ldots, X_n)$ with $\nu = \mu$ and $\kappa = k\rho$.

We shall consider subsets $\mathscr{F} \subset \mathbb{L}^2(A, \nu\otimes\mu)$ corresponding to smoothness or structural assumptions on $\sqrt{s}|_A$. For such an $\mathscr{F}$, we associate a collection $\mathbb{V}$ and deduce from Theorem 5 a risk bound for the estimator $\hat{s}$ when $\sqrt{s}|_A$ belongs to $\mathscr{F}$. This set is a generic notation and will change from section to section. In the remaining part of this chapter, we shall always choose $\mathbb{X}^2 = \mathbb{R}^{2d}$, $A = [0,1]^{2d}$ and μ the Lebesgue measure.

4.3. Smoothness assumptions. We have introduced in Section 2.6 the isotropic Besov spaces $\mathscr{B}_q^\sigma(\mathbb{L}^p([0,1]^{2d}))$ where $\sigma \in (0,1)$. In this section, we consider the anisotropic Besov spaces $\mathscr{B}_q^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2d})$ belongs to $(0,+\infty)^{2d}$.

Intuitively, a function f on $[0,1]^{2d}$ belongs to $\mathscr{B}_q^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ if, for all $j \in \{1, \ldots, 2d\}$, and $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{2d} \in [0,1]$ the function

$$x_j \mapsto f(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_{2d})$$

belongs to $\mathscr{B}_q^{\sigma_j}(\mathbb{L}^p([0,1]))$. In particular, for all $\sigma \in (0,+\infty)$,

$$\mathscr{B}_q^\sigma(\mathbb{L}^p([0,1]^{2d})) = \mathscr{B}_q^{(\sigma,\ldots,\sigma)}(\mathbb{L}^p([0,1]^{2d})).$$

A definition of the anisotropic Besov spaces may be found in Hochmuth (2002) (for $d = 1$) and in Akakpo (2009) (for larger values of d). We also consider the space $\mathcal{H}^{\boldsymbol{\sigma}}([0,1]^{2d})$ of anisotropic

Hölderian functions on $[0,1]^{2d}$ with regularity $\boldsymbol{\sigma}$. A precise definition of this space may be found in Section 3.1.1 of Baraud and Birgé (2011) (among other references).

For all $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2d}) \in (0,+\infty)^{2d}$, we denote by $\bar{\boldsymbol{\sigma}}$ the harmonic mean of $\boldsymbol{\sigma}$:

$$\frac{1}{\bar{\boldsymbol{\sigma}}} = \frac{1}{2d} \sum_{i=1}^{2d} \frac{1}{\sigma_i}.$$

We set for all $p \in (0,+\infty]$,

$$\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d})) = \begin{cases} \mathscr{B}^{\boldsymbol{\sigma}}_{\infty}(\mathbb{L}^p([0,1]^{2d})) & \text{if } p \in (0,1] \\ \mathscr{B}^{\boldsymbol{\sigma}}_{p}(\mathbb{L}^p([0,1]^{2d})) & \text{if } p \in (1,2) \\ \mathscr{B}^{\boldsymbol{\sigma}}_{\infty}(\mathbb{L}^p([0,1]^{2d})) & \text{if } p \in [2,+\infty) \\ \mathcal{H}^{\boldsymbol{\sigma}}([0,1]^{2d}) & \text{if } p = \infty \end{cases}$$

and denote by $|\cdot|_{p,\boldsymbol{\sigma}}$ the semi norm associated to the space $\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$.

In this section, we are interesting in obtaining a bound risk when $\sqrt{s}|_A$ belongs to the space

$$\mathscr{B}([0,1]^{2d}) = \bigcup_{p \in (0,+\infty]} \left(\bigcup_{\substack{\boldsymbol{\sigma} \in (0,+\infty)^{2d} \\ \bar{\boldsymbol{\sigma}} > 2d(1/p-1/2)_+}} \mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d})) \right).$$

Families of linear spaces possessing good approximation properties with respect to the elements of $\mathscr{F} = \mathscr{B}([0,1]^{2d})$ can be found in Theorem 1 of Akakpo (2012). We then deduce from Theorem 5,

Corollary 3. *Suppose that Assumption 3 holds with $\mathbb{X} = \mathbb{R}^d$, $A = [0,1]^{2d}$ and with $\nu \otimes \mu$ the Lebesgue measure. There exists an estimator $\hat{s}$ such that for all $\sqrt{s}|_A \in \mathscr{B}([0,1]^{2d})$,*

$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \left|\sqrt{s}\right|_A\Big|^{2d/(d+\bar{\boldsymbol{\sigma}})}_{p,\boldsymbol{\sigma}} \left(\frac{\log n}{n}\right)^{\bar{\boldsymbol{\sigma}}/(\bar{\boldsymbol{\sigma}}+d)} + \frac{\log n}{n}$$

where $p \in (0,+\infty]$, $\boldsymbol{\sigma} \in (0,+\infty)^{2d}$, $\bar{\boldsymbol{\sigma}} > 2d(1/p - 1/2)_+$ are such that $\sqrt{s}|_A \in \mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ and where $C > 0$ depends only on κ,d,p,$\boldsymbol{\sigma}$.

To our knowledge, the only statistical procedures that can adapt both to possible inhomogeneity and anisotropy of s are those of Akakpo and Lacour (2011) and Birgé (2012). The losses are different, but the rates are the same as ours (up to the logarithmic term). In view of our assumptions, we do not know if the logarithmic term can be avoided.

In the following sections, we consider classes $\mathscr{F}$ corresponding to structural assumptions on $\sqrt{s}|_A$. More precisely, rates of convergence when the chain is autoregressive with constant conditional variance (respectively non constant conditional variance) are established in Section 4.4 (respectively Section 4.5).

4.4. AR model. In this section, we assume that $X_{n+1} = g(X_n) + \varepsilon_n$ where g is an unknown function and where the ε_n's are unobserved identically distributed random variables. Many papers are devoted to the estimation of the regression function g and it is beyond the scope of this work to make an historical review for this statistical problem.

For the sake of simplicity, one shall assume throughout this section that $\mathbb{X} = \mathbb{R}$, $A = [0,1]^2$. The transition density is of the form $s(x,y) = \varphi(y - g(x))$ where φ is the density of ε_0. Since g and φ are both unknown, this suggests us to consider the class

$$\mathscr{F} = \bigcup_{\sigma>0} \left\{ f,\, \exists \phi \in \mathcal{H}^\sigma(\mathbb{R}), \exists g \in \mathscr{B}([0,1]), \|g\|_\infty < \infty,\, \forall x,y \in [0,1],\ f(x,y) = \phi(y - g(x)) \right\}.$$

A family $\mathbb{V}$ of linear spaces possessing good approximation properties with respect to the functions of $\mathscr{F}$ can be built by using Section 6.2 of Baraud and Birgé (2011). Precisely, we prove the following.

Corollary 4. *Suppose that Assumption 3 holds with $\mathbb{X} = \mathbb{R}$, $A = [0,1]^2$ and with $\nu \otimes \mu$ the Lebesgue measure on $\mathbb{R}^2$. Assume that $\sqrt{s}|_A$ belongs to $\mathscr{F}$. Let $\sigma > 0$, $p \in (0, +\infty]$, $\beta > (1/p - 1/2)_+$ be any numbers and $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $g \in \mathscr{B}^\beta(\mathbb{L}^p([0,1]))$, $\|g\|_\infty < \infty$ be any functions such that*

$$\sqrt{s(x,y)} = \phi(y - g(x)) \quad \textit{for all } x,y \in [0,1].$$

There exists two estimators $\hat{\phi} \geq 0$ and $\hat{g}$ such that the estimator $\hat{s}$ defined by

$$\hat{s}(x,y) = \left(\hat{\phi}\,(y - \hat{g}(x))\right)^2 \mathbb{1}_{[0,1]^2}(x,y) \quad \textit{for all } x,y \in \mathbb{R}$$

satisfies

$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq C_1' \left(\frac{\log^2 n}{n}\right)^{\frac{2\beta(\sigma\wedge 1)}{2\beta(\sigma\wedge 1)+1}} + C_2' \left(\frac{\log n}{n}\right)^{\frac{2\sigma}{2\sigma+1}}$$

where $C > 0$ depends only on κ,p,σ,β, where C_1' depends only on p,β,σ,$|g|_{p,\beta}$,$\|g\|_\infty$,$|\phi|_{\infty,\sigma\wedge 1}$ and where C_2' depends only on σ,$\|g\|_\infty$,$|\phi|_{\infty,\sigma}$. Moreover, the construction of the estimators $\hat{g}$, $\hat{\phi}$ depends only on the data $X_0, \ldots, X_n$.

In particular, if ϕ is very smooth (says $\sigma \geq \beta \vee 1$), the rate of convergence corresponds to the rate of convergence for estimating g only (up to a logarithmic term).

It is interesting to compare the preceding rate to the one we would obtain under the pure smoothness assumption on $\sqrt{s}|_A$ but ignoring that $\sqrt{s}|_A$ belongs to $\mathscr{F}$. To do so, we need to specify the regularity of $\sqrt{s}|_A$, knowing that of ϕ and g. This is the purpose of the following lemma.

Lemma 1. *Let $\sigma, \beta > 0$, and let us define*

$$\theta(\beta,\sigma) = \begin{cases} \beta\sigma & \textit{if } \beta,\sigma \leq 1 \\ \beta \wedge \sigma & \textit{otherwise.} \end{cases}$$

Let $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $g \in \mathcal{H}^\beta([0,1])$. The function f defined by

$$f(x,y) = \phi(y - g(x)) \quad \textit{for all } x,y \in [0,1],$$

belongs to $\mathcal{H}^{(\theta(\beta,\sigma),\sigma)}([0,1]^2)$.

Moreover, for all $\sigma,\beta>0$, *there exist* $\phi\in\mathcal{H}^\sigma(\mathbb{R})$, $g\in\mathcal{H}^\beta([0,1])$ *such that the function* f *defined by*

$$f(x,y)=\phi(y-g(x)) \quad \textit{for all } x,y\in[0,1],$$

belongs to $\mathcal{H}^{(a,b)}([0,1]^2)$ *if and only if* $a\le\theta(\beta,\sigma)$ *and* $b\le\sigma$.

This result says that if $\sqrt{s(x,y)}=\phi(y-g(x))$, with $\phi\in\mathcal{H}^\sigma(\mathbb{R})$, $g\in\mathcal{H}^\beta([0,1])$, then $\sqrt{s}$ is Hölderian with regularity $(\theta(\beta,\sigma),\sigma)$ on $[0,1]^2$, and this regularity cannot be improved in general except in some particular situations. Under such a smoothness assumption, the rate of estimation we would get is $(\log n/n)^{2\sigma\theta(\beta,\sigma)/(2\sigma\theta(\beta,\sigma)+\theta(\beta,\sigma)+\sigma)}$. This rate is always slower than the rate obtained under the structural assumption.

4.5. ARCH model. Throughout this section, we assume that $X_{n+1}=g_1(X_n)+g_2(X_n)\varepsilon_n$ where g_1,g_2 are unknown functions and where the ε_n's are unobserved identically distributed random variables. The previous model corresponded to $g_2=1$. The problem of the estimation of the mean and variance functions g_1 and g_2 was considered in several papers and we refer to Section 1.2 of Comte and Rozenholc (2002) for bibliographical references.

For the sake of simplicity, one assumes that $\mathbb{X}=\mathbb{R}$ and $A=[0,1]^2$. If φ denotes the density of ε_0, the transition density s is of the form

$$s(x,y)=|g_2(x)|^{-1}\varphi\left[g_2^{-1}(x)\,(y-g_1(x))\right] \quad \text{for all } x,y\in\mathbb{R}. \tag{7}$$

We consider thus the class

$$\begin{aligned}\mathscr{F} \;=\; \bigcup_{\sigma>0}\Big\{&f,\ \exists\phi\in\mathcal{H}^\sigma(\mathbb{R}),\exists v_1,v_2\in\mathscr{B}([0,1]),\ \|v_1\|_\infty<\infty,\ \|v_2\|_\infty<\infty,\\ &\forall x,y\in[0,1],\ f(x,y)=\sqrt{|v_2(x)|}\phi\left(v_2(x)(y-v_1(x))\right)\Big\}\end{aligned}$$

and apply Theorem 5 with a suitable collection $\mathbb{V}$ to obtain:

Corollary 5. *Suppose that Assumption 3 holds with* $\mathbb{X}=\mathbb{R}$, $A=[0,1]^2$ *and with* $\nu\otimes\mu$ *the Lebesgue measure on* $\mathbb{R}^2$. *Assume that* $\sqrt{s}|_A$ *belongs to* $\mathscr{F}$. *Let* $\sigma>0$, $\phi\in\mathscr{B}^\sigma(\mathbb{R})$ *and for all* $i\in\{1,2\}$, *let* $p_i\in(0,+\infty]$, $\beta_i>(1/p_i-1/2)_+$, $v_i\in\mathscr{B}^{\beta_i}(\mathbb{L}^{p_i}([0,1]))$, *with* $\|v_i\|_\infty<\infty$ *such that*

$$\sqrt{s(x,y)}=\sqrt{|v_2(x)|}\phi\left(v_2(x)(y-v_1(x))\right) \quad \textit{for all } x,y\in[0,1].$$

Let $p_3\in(0,+\infty]$ *and* $\beta_3>(1/p_3-1/2)_+$ *be any numbers such that* $v_3=\sqrt{|v_2|}\in\mathscr{B}^{\beta_3}(\mathbb{L}^{p_3}([0,1]))$. *There exists an estimator* $\hat{s}$ *such that*

$$C\mathbb{E}\left[H^2\left(s\mathbb{1}_A,\hat{s}\right)\right]\le C_1'\left(\frac{\log^2 n}{n}\right)^{\frac{2\beta(\sigma\wedge1)}{2\beta(\sigma\wedge1)+1}}+C_2'\left(\frac{\log n}{n}\right)^{\frac{2\sigma}{2\sigma+1}}$$

where $\beta=\max(\beta_1,\beta_2,\beta_3)$. *The constant* $C>0$ *depends only on* κ,σ,p_1,p_2,p_3,β_1,β_2,β_3, C_1' *depends only on* σ,$\|v_1\|_\infty$,$\|v_2\|_\infty$,$\|\varphi\|_\infty$,$|v_1|_{p_1,\beta_1}$,$|v_2|_{p_2,\beta_2}$,$|v_3|_{p_3,\beta_3}$,$|\varphi|_{\infty,\sigma\wedge1}$ *and* C_2' *depends only on* σ,$\|v_2\|_\infty$,$|\varphi|_{\infty,\sigma}$. *Moreover, the construction of the estimator* $\hat{s}$ *depends only on the data* $X_0,\dots,X_n$.

If s is of the form (7) with φ, g_1, g_2 smooth, in the sense that $\phi = \sqrt{\varphi} \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 = g_1 \in \mathscr{B}^{\beta_1}(\mathbb{L}^{p_1}([0,1]))$, $\|v_1\|_\infty < \infty$, $v_2 = g_2^{-1} \in \mathscr{B}^{\beta_2}(\mathbb{L}^{p_2}([0,1]))$, $\|v_2\|_\infty < \infty$ and $v_3 = |g_2|^{-1/2} \in \mathscr{B}^{\beta_3}(\mathbb{L}^{p_3}([0,1]))$, then $\sqrt{s}|_A$ belongs to $\mathscr{F}$. If ϕ is sufficiently smooth ($\sigma \geq \beta_1 \vee \beta_2 \vee \beta_3 \vee 1$), the rate becomes

$$C''\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \max\left(\left(\frac{\log^2 n}{n}\right)^{\frac{2\beta_1}{2\beta_1+1}}, \left(\frac{\log^2 n}{n}\right)^{\frac{2\beta_2}{2\beta_2+1}}, \left(\frac{\log^2 n}{n}\right)^{\frac{2\beta_3}{2\beta_3+1}}\right).$$

Up to a logarithmic term, the first term corresponds to the bound we would get if we could estimate g_1 only. The two other terms correspond to the rate of estimation of g_2^{-1} and $|g_2|^{-1/2}$ respectively (up to a logarithmic term).

Note that if $\beta_2 \in (0,1)$, one can always choose $p_3 = 2p_2$ (with $p_3 = \infty$ if $p_2 = \infty$), $\beta_3 = \beta_2/2$, in which case the rate becomes

$$C''\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \max\left(\left(\frac{\log^2 n}{n}\right)^{\frac{2\beta_1}{2\beta_1+1}}, \left(\frac{\log^2 n}{n}\right)^{\frac{\beta_2}{\beta_2+1}}\right).$$

In some situations however, β_3 can be taken larger than β_2.

As in the preceding section, we may use the lemma below to compare this rate with the one we would obtain under smoothness assumptions on $\sqrt{s}|_A$.

Lemma 2. *Let for all $\sigma, \beta_1, \beta_2 > 0$,*

$$\theta(\beta_1, \beta_2, \sigma) = \begin{cases} \left(2^{-1}(\beta_2 \wedge 1)\right) \wedge \sigma\beta_1 \wedge \sigma\beta_2 & \text{if } \sigma \leq 1 \text{ and } \beta_1 \wedge \beta_2 \leq 1 \\ \left(2^{-1}(\beta_2 \wedge 1)\right) \wedge \sigma \wedge \beta_1 & \text{otherwise.} \end{cases}$$

Let $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$. The function f defined by

$$f(x,y) = \sqrt{|v_2(x)|}\phi\left(v_2(x)(y - v_1(x))\right) \quad \text{for all } x, y \in [0,1],$$

belongs to $\mathcal{H}^{(\theta(\beta_1,\beta_2,\sigma),\sigma)}([0,1]^2)$.

Moreover, there exist $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$ such that the function f defined by

$$f(x,y) = \sqrt{|v_2(x)|}\phi\left(v_2(x)(y - v_1(x))\right) \quad \text{for all } x, y \in [0,1],$$

belongs to $\mathcal{H}^{(a,b)}([0,1]^2)$ if and only if $a \leq \theta(\beta_1, \beta_2, \sigma)$ and $b \leq \sigma$.

This proposition says that if $\sqrt{s(x,y)} = \sqrt{|v_2(x)|}\phi\left(v_2(x)(y - v_1(x))\right)$, with $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$, $\sqrt{s}|_A$ belongs to $\mathcal{H}^{(\theta(\beta_1,\beta_2,\sigma),\sigma)}([0,1]^2)$ and the regularity index of this space cannot be increased in general. By Corollary 3, we would get a rate of order $(\log n/n)^{2\theta(\beta_1,\beta_2,\sigma)\sigma/(2\theta(\beta_1,\beta_2,\sigma)\sigma+\theta(\beta_1,\beta_2,\sigma)+\sigma)}$, which is slower than the one given by Corollary 5.

5. Appendix: implementation of the first procedure

In this section, we explain how to construct in practice the estimator of the first procedure. This will lead to the proposition below.

Proposition 6. *For all $L > 0$, $\ell \in \mathbb{N}^\star$, the estimator $\hat{s} = \hat{s}(L, \ell)$ of Section 2.3 can be built in less than $C\left(n\ell d + \ell 4^{(\ell+1)d}\right)$ operations where C is an universal constant.*

We set for all $K \in \cup_{m\in\mathcal{M}_\ell} m$,

$$\hat{s}_K = \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x)\, \mathrm{d}\mu(x)} \mathbb{1}_K,$$

for all $K' \in \cup_{m\in\mathcal{M}_\ell} m$,

$$F_K(K') = \alpha H^2\left(\hat{s}_K \mathbb{1}_{K'}, \hat{s}_{K'} \mathbb{1}_K\right) + T\left(\hat{s}_K \mathbb{1}_{K'}, \hat{s}_{K'} \mathbb{1}_K\right),$$

and for all $m' \in \mathcal{M}_\ell$,

$$\gamma_K(m') = \left(\sum_{K'\in m'} F_K(K')\right) - \mathrm{pen}(m' \vee K).$$

We shall find for each cube $K \in \cup_{m\in\mathcal{M}_\ell} m$, a partition $m'_K \in \mathcal{M}_\ell$ such that

$$\gamma_K(m'_K) = \sup_{m'\in\mathcal{M}_\ell} \gamma_K(m'). \tag{8}$$

We shall compute then

$$\min_{m\in\mathcal{M}_\ell} \gamma(m) = \min_{m\in\mathcal{M}_\ell} \left\{ \left(\sum_{K\in m} \gamma_K(m'_K)\right) + 2\mathrm{pen}(m) \right\}. \tag{9}$$

We shall find m'_K by using a slight adaptation of the procedure of Blanchard et al. (2004). Computing (9) is similar. The algorithm we propose is based on the one-to-one correspondence between $\mathcal{M}_\ell$ and the set $\mathcal{T}_\ell$ of 4^d-ary trees with depth smaller than ℓ.

Lemma 3. *There exists a one-to-one map ψ_ℓ between $\mathcal{M}_\ell$ and $\mathcal{T}_\ell$ such that for all $m \in \mathcal{M}_\ell$, $\psi_\ell(m)$ is a tree whose leaves correspond to the elements of the partition m.*

The construction of this map may for instance be deduced from Section 3.2.4 of Baraud and Birgé (2009).

We need to introduce some notations. For each tree $T \in \mathcal{T}_\ell$ and bin K'' of T, we denote by $T(K'')$ the subtree of T rooted in K''. The set of leaves of $T(K'')$ is denoted by $\mathcal{L}(T(K''))$. We set $R(K'')$ the tree reduced to its root K'' (*i.e*, $\mathcal{L}(R(K'')) = \{K''\}$). For all cube $K \in \cup_{m\in\mathcal{M}_\ell} m$, we set

$$\mathcal{L}(T(K'')) \vee K = \left\{K' \cap K,\, K' \in \mathcal{L}(T(K'')),\, K' \cap K \neq \emptyset\right\}$$

and we define the function $\mathcal{E}$ by

$$\mathcal{E}(T(K'')) = -|\mathcal{L}(T(K'')) \vee K| + \sum_{K'\in\mathcal{L}(T(K''))} F_K(K').$$

The key point is that computing (8) amounts to finding $T^\star$ such that

$$\mathcal{E}(T^\star([0,1]^{2d})) = \sup_{T\in\mathcal{T}_\ell} \mathcal{E}(T([0,1]^{2d}))$$

since $m'_K = \psi_\ell^{-1}(T^\star)$.

We now take advantage of the additivity of the function $\mathcal{E}$: if $T(K'')$ is not reduced to its root, and if $K''_1, \dots, K''_{4^d}$ are the cubes of $\cup_{m\in\mathcal{M}_\ell} m$ such that $K'' = \cup_{i=1}^{4^d} K''_i$, then,

$$\mathcal{E}(T(K'')) = \sum_{i=1}^{4^d} \mathcal{E}(T(K''_i)). \tag{10}$$

For all cube $K'' \in \cup_{m\in\mathcal{M}_\ell} m$, let $T^\star(K'')$ be a tree (rooted in K'') such that

$$\mathcal{E}(T^\star(K'')) = \sup_{T\in\mathcal{T}_\ell, T\ni K''} \mathcal{E}(T(K'')).$$

Remark that if $K'' \cap K = \emptyset$, this supremum is equal to 0, in which case $T^\star(K'')$ will always stand for $R(K'')$. In general, we deduce from (10) that

$$\mathcal{E}(T^\star(K'')) = \max\left(\mathcal{E}(R(K'')), \sum_{i=1}^{4^d} \mathcal{E}(T^\star(K''_i))\right). \tag{11}$$

Calculating (8) can thus be completed in that way: we start with the sets $K'' \in \cup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_{\ell-1}} m$ with $K'' \cap K \neq \emptyset$ for which the optimal local trees are reduced to their roots. By using relation (11) we find the optimal local trees $T^\star(K'')$ when $K'' \in \cup_{\mathcal{M}_{\ell-1}\setminus\mathcal{M}_\ell} m$, $K'' \cap K \neq \emptyset$. Proceeding recursively like this leads to the optimal tree $T^\star = T^\star([0,1]^{2d})$.

6. Proofs

6.1. Proof of Proposition 1. Let us introduce the piecewise constant function

$$\bar{s}_m = \sum_{K\in m} \frac{\sum_{i=0}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_K. \tag{12}$$

By using the triangular inequality we can decompose the risk of $\hat{s}_m$ as follows:

$$\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s}_m)\right] \leq \left(1 + \frac{2+\log 2}{2}\right) \mathbb{E}\left[H^2(s\mathbb{1}_A, \bar{s}_m)\right] + \left(1 + \frac{2}{2+\log 2}\right) \mathbb{E}\left[H^2(\bar{s}_m, \hat{s}_m)\right].$$

The first term can be bounded from above by $(4+\log 2)\,\mathbb{E}\left[H^2(s\mathbb{1}_A, V_m)\right]$ thanks to Lemma 2 of Baraud and Birgé (2009). For the second term, we begin to define for $K \in m$ the random variable

$$B_K = \left(\sqrt{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\right)^2.$$

Since $2nH^2(\hat{s}_m, \bar{s}_m) = \sum_{K\in m} B_K$, we shall bound from above the terms $\mathbb{E}[B_K]$. For this purpose, we introduce the stopping time

$$T = \inf\left\{i \geq 0,\ \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right] \geq \frac{1}{2n}\right\} \wedge (n-1)$$

with respect to the filtration $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ generated by the random variables $X_0, \dots, X_n$. We set $\varepsilon = 1 + \log 2 + 2\log n$ and use the algebraic inequality

$$\left(\sqrt{a+b} - \sqrt{c+d}\right)^2 \leq (1+\varepsilon)\left(\sqrt{a} - \sqrt{c}\right)^2 + (1+\varepsilon^{-1})\left(\sqrt{b} - \sqrt{d}\right)^2$$

to decompose $\mathbb{E}[B_K]$:

$$\begin{aligned}\mathbb{E}\left[B_K\right] &\leq (1+\varepsilon)\,\mathbb{E}\left[\left(\sqrt{\sum_{i=0}^{T-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{T-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\right)^2\right] \\ &\quad + \left(1+\varepsilon^{-1}\right)\mathbb{E}\left[\left(\sqrt{\sum_{i=T}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=T}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\right)^2\right].\end{aligned}$$

By using $\left(\sqrt{a} - \sqrt{b}\right)^2 \leq (a-b)^2/b$,

$$\begin{aligned}\mathbb{E}\left[B_K\right] &\leq 2\,(1+\varepsilon)\,\mathbb{E}\left[\sum_{i=0}^{T-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right] \\ &\quad + \left(1+\varepsilon^{-1}\right)\mathbb{E}\left[\frac{\left(\sum_{i=T}^{n-1}\left(\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right)\right)^2}{\sum_{i=T}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\right]. \quad (13)\end{aligned}$$

Yet,

$$\mathbb{E}\left[\sum_{i=0}^{T-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right] \leq 1/2,$$

and we control the second term of the right-hand side of inequality (13), thanks to the claims below.

Claim 1. *For all $K \in m$, $j \in \{0, \dots, n\}$, and $\mathcal{A}' \in \mathcal{F}_j = \sigma(X_0, \dots, X_j)$,*

$$\mathbb{E}\left[\frac{\left(\sum_{i=j}^{n-1}\left(\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right)\right)^2}{\sum_{i=j}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]} \mathbb{1}_{\mathcal{A}'}\right] \leq \sum_{k=j}^{n-1} \mathbb{E}\left[\frac{\mathbb{E}\left[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k\right]}{\sum_{i=j}^{k} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]} \mathbb{1}_{\mathcal{A}'}\right].$$

Proof of Claim 1. Let us define the random variables

$$Y_{n-1}(K) = \sum_{i=j}^{n-1}\left(\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right) \text{ and } Z_n(K) = \frac{Y_{n-1}^2(K)}{\sum_{i=j}^{n-1} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}.$$

We have

$$\begin{aligned}\mathbb{E}\left[Z_{n+1}(K) \mid \mathcal{F}_n\right] &= \frac{\mathbb{E}\left(\left[Y_{n-1}(K) + \left(\mathbb{1}_K(X_n, X_{n+1}) - \mathbb{E}\left[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n\right]\right)\right]^2 \mid \mathcal{F}_n\right)}{\sum_{i=j}^{n} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]} \\ &= \frac{Y_{n-1}^2(K) + \operatorname{var}\left(\mathbb{1}_K(X_n, X_{n+1}) \mid X_n\right)}{\sum_{i=j}^{n} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}.\end{aligned}$$

Hence,

$$\mathbb{E}\left[Z_{n+1}(K) \mid \mathcal{F}_n\right] \leq Z_n(K) + \frac{\mathbb{E}\left[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n\right]}{\sum_{i=j}^n \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}$$

and, since $\mathcal{A}'$ is also $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[Z_{n+1}(K)\mathbb{1}_{\mathcal{A}'}\right] \leq \mathbb{E}\left[Z_n(K)\mathbb{1}_{\mathcal{A}'}\right] + \mathbb{E}\left[\frac{\mathbb{E}\left[\mathbb{1}_K(X_n, X_{n+1}) \mid X_n\right]}{\sum_{i=j}^n \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\mathbb{1}_{\mathcal{A}'}\right].$$

The result ensues from induction. □

Claim 2. *For all sequence $(u_n)_{n\geq 0}$ in $[0,1]$, and $j \geq 0$ such that $u_j \neq 0$,*

$$\sum_{k=j}^{n-1} \frac{u_k}{\sum_{i=j}^k u_i} \leq 1 + \log n - \log u_j.$$

Proof of Claim 2. Let f be any non-negative continuous function such that $u_k = \int_k^{k+1} f(t)\,\mathrm{d}t$ whatever $k \in \mathbb{N}$. Let F be the primitive of f such that $F(j) = 0$. Then,

$$\begin{aligned}
\sum_{k=j}^{n-1} \frac{u_k}{\sum_{i=j}^k u_i} &\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(k+1)}\,\mathrm{d}t \\
&\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(t)}\,\mathrm{d}t \\
&\leq 1 + \log F(n) - \log F(j+1) \\
&\leq 1 + \log\left(\sum_{k=j}^{n-1} u_k\right) - \log u_j.
\end{aligned}$$

□

By using Claim 1 with $\mathcal{A}' = [T = j]$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\sum_{i=T}^{n-1}\left(\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]\right)\right)^2}{\sum_{i=T}^{n-1}\mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\right] &\leq \sum_{j=0}^{n-2}\mathbb{E}\left(\sum_{k=j}^{n-1}\frac{\mathbb{E}\left[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k\right]}{\sum_{i=j}^k \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]}\mathbb{1}_{T=j}\right) \\
&\quad + \mathbb{E}\left[\frac{\left(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}\left[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right]\right)^2}{\mathbb{E}\left[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right]}\mathbb{1}_{T=n-1}\right].
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}\left[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right]\right)^2}{\mathbb{E}\left[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right]}\mathbb{1}_{T=n-1}\right] &= \mathbb{E}\left[\frac{\mathrm{var}\left(\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right)}{\mathbb{E}\left[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}\right]}\mathbb{1}_{T=n-1}\right] \\
&\leq \mathbb{P}\left(T = n-1\right).
\end{aligned}$$

We then use Claim 2 with $u_k = \mathbb{E}\left[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k\right]$ to derive

$$\begin{aligned}\sum_{j=0}^{n-2} \mathbb{E}\left(\sum_{k=j}^{n-1} \frac{\mathbb{E}\left[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k\right]}{\sum_{i=j}^{k} \mathbb{E}\left[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\right]} \mathbb{1}_{T=j}\right) &\leq \sum_{j=0}^{n-2} \mathbb{E}\left[(1+\log 2 + 2\log n)\mathbb{1}_{T=j}\right] \\ &\leq (1+\log 2 + 2\log n)\,\mathbb{P}\left(T \neq n-1\right).\end{aligned}$$

Finally, $\mathbb{E}\left[B_K\right] \leq 4 + 2\log 2 + 4\log n$ and hence

$$\mathbb{E}\left[H^2(\bar{s}_m, \hat{s}_m)\right] \leq \frac{2+\log 2 + 2\log n}{n}|m|,$$

which concludes the proof. □

6.2. Proof of Theorem 2. When $\ell \leq n$, the result ensues from the following theorem whose proof is delayed to Section 6.3. In the theorem below, the constant $L_0 = 90$ can easily be improved but it seems to be difficult to obtain the value $L_0 = 0.03$ used in practice.

Theorem 7. *For all $L \geq 90$ and $1 \leq \ell \leq n$, the estimator $\hat{s} = \hat{s}(L, \ell)$ satisfies*

$$\forall \xi > 0, \quad \mathbb{P}\left[CH^2\left(s\mathbb{1}_A, \hat{s}\right) \geq \inf_{m \in \mathcal{M}_\ell} \left(H^2\left(s\mathbb{1}_A, \hat{s}_m\right) + \mathrm{pen}(m)\right) + \xi\right] \leq 3e^{-n\xi}$$

where C is an universal positive constant.

By integrating the inequality above, there exists $C' > 0$ such that

$$C'\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \inf_{m \in \mathcal{M}_\ell} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}_m\right)\right] + \mathrm{pen}(m)\right\}$$

and the conclusion follows from Proposition 1.

When ℓ is larger than n, we use the lemma below whose proof is postponed to Section 6.3.3.

Lemma 4. *For all $L \geq 15$ and $\ell \geq n+1$, $\hat{s}(L, \ell) = \hat{s}(L, n)$ and $\hat{s}(L, \infty) = \hat{s}(L, n)$.*

For $L \geq 90$, if $\ell \geq n+1$ or $\ell = \infty$, we have thus

$$C'\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \hat{s}\right)\right] \leq \inf_{m \in \mathcal{M}_n} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] + \mathrm{pen}(m)\right\}.$$

Let $m^\star \in \mathcal{M}_\ell$ such that

$$2 \inf_{m \in \mathcal{M}_\ell} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] + \mathrm{pen}(m)\right\} \geq \mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_{m^\star}\right)\right] + \mathrm{pen}(m^\star).$$

Since $\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] \leq \mathbb{E}\left[H^2\left(s\mathbb{1}_A, 0\right)\right] \leq 1/2$ and $\{[0,1]^{2d}\} \in \mathcal{M}_\ell$,

$$\inf_{m \in \mathcal{M}_\ell} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] + \mathrm{pen}(m)\right\} \leq \frac{1}{2} + L\frac{\log n}{n}.$$

Consequently, $L|m^\star|\log(n)/n \leq 1 + 2L\log(n)/n$ and thus $|m^\star| \leq 2 + n/(L\log n) \leq n$. Remark now that the cardinality of a partition $m \in \mathcal{M}_\ell \setminus \mathcal{M}_n$ can be lower bounded by

$$|m| \geq 4^d + (4^d - 1)n \geq n + 1.$$

Consequently, $m^\star \in \mathcal{M}_n$ and hence,

$$\inf_{m \in \mathcal{M}_n} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] + \mathrm{pen}(m)\right\} \leq 2 \inf_{m \in \mathcal{M}_\ell} \left\{\mathbb{E}\left[H^2\left(s\mathbb{1}_A, V_m\right)\right] + \mathrm{pen}(m)\right\}$$

which completes the proof. □

6.3. Proof of Theorem 7. The proof of this theorem requires the two following lemmas whose proofs are postponed to Sections 6.3.1 and 6.3.2.

Lemma 5. *For all partition $m \in \mathcal{M}_\ell$,*

$$\gamma_1(m) = \sup_{m' \in \mathcal{M}_\ell} \left\{ \alpha H^2 \left(\hat{s}_m, \hat{s}_{m'} \right) + T \left(\hat{s}_m, \hat{s}_{m'} \right) - \mathrm{pen}(m') \right\} + \mathrm{pen}(m)$$

satisfies $\gamma_1(m) \leq \gamma(m)$.

For all $m \in \mathcal{M}_\ell$, there exists a deterministic set S_m such that $\hat{s}_m \in S_m$ and such that

$$\gamma_2(m) = \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{ \alpha H^2 \left(\hat{s}_m, f' \right) + T \left(\hat{s}_m, f' \right) - \mathrm{pen}(m') \right\} + 2\mathrm{pen}(m)$$

satisfies $\gamma(m) \leq \gamma_2(m)$.

Lemma 6. *Set $\varepsilon = \left(2 + 3\sqrt{2}\right)/8$. Under the assumptions of Theorem 7, for all $\xi > 0$, there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - 3e^{-n\xi}$ and on which,*

$$\text{for all partition } m \in \mathcal{M}_\ell, \qquad (14)$$
$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{ (1-\varepsilon) H^2 \left(s\mathbb{1}_A, f' \right) + T \left(\hat{s}_m, f' \right) - \mathrm{pen}(m') \right\} \leq (1+\varepsilon) H^2 \left(s\mathbb{1}_A, \hat{s}_m \right) + \mathrm{pen}(m) + 22\xi$$

where $S_{m'}$ is defined in Lemma 5.

Proof of Theorem 7. On Ω_ξ, for all $m \in \mathcal{M}_\ell$,

$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{ (1-\varepsilon) H^2 \left(s\mathbb{1}_A, f' \right) + T \left(\hat{s}_m, f' \right) - \mathrm{pen}(m') \right\} \leq (1+\varepsilon) H^2 \left(s\mathbb{1}_A, \hat{s}_m \right) + \mathrm{pen}(m) + 22\xi.$$

If $T \left(\hat{s}_m, \hat{s}_{\hat{m}} \right) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \geq 0$,

$$\begin{aligned} \alpha H^2 \left(s\mathbb{1}_A, \hat{s}_{\hat{m}} \right) &\leq (1-\varepsilon) H^2 \left(s\mathbb{1}_A, \hat{s}_{\hat{m}} \right) + T \left(\hat{s}_m, \hat{s}_{\hat{m}} \right) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m) \\ &\leq (1+\varepsilon) H^2 \left(s\mathbb{1}_A, \hat{s}_m \right) + 2\mathrm{pen}(m) + 22\xi \end{aligned}$$

since $\alpha \leq 1 - \varepsilon$ and since $\hat{s}_{\hat{m}}$ belongs to $\cup_{m' \in \mathcal{M}_\ell} S_{m'}$.

If $T \left(\hat{s}_m, \hat{s}_{\hat{m}} \right) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) < 0$,

$$\alpha H^2 \left(\hat{s}_m, \hat{s}_{\hat{m}} \right) \leq \alpha H^2 \left(\hat{s}_{\hat{m}}, \hat{s}_m \right) + T \left(\hat{s}_{\hat{m}}, \hat{s}_m \right) - \mathrm{pen}(m) + \mathrm{pen}(\hat{m})$$

since $T \left(\hat{s}_{\hat{m}}, \hat{s}_m \right) = -T \left(\hat{s}_m, \hat{s}_{\hat{m}} \right)$. Hence,

$$\begin{aligned} \alpha H^2 \left(\hat{s}_m, \hat{s}_{\hat{m}} \right) &\leq \sup_{m' \in \mathcal{M}_\ell} \left\{ \alpha H^2 \left(\hat{s}_{\hat{m}}, \hat{s}_{m'} \right) + T \left(\hat{s}_{\hat{m}}, \hat{s}_{m'} \right) - \mathrm{pen}(m') \right\} + \mathrm{pen}(\hat{m}) \\ &\leq \gamma_1(\hat{m}). \end{aligned}$$

By using Lemma 5,

$$\gamma_1(\hat{m}) \leq \gamma(m) + \frac{1}{n} \leq \gamma_2(m) + \frac{1}{n},$$

and thus

$$\alpha H^2\left(\hat{s}_m, \hat{s}_{\hat{m}}\right) \leq \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{\alpha H^2\left(\hat{s}_m, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}(m')\right\} + 2\mathrm{pen}(m) + \frac{1}{n}.$$

With $\upsilon = (1-\varepsilon)/\alpha - 1 > 0$,

$$\begin{aligned}
\alpha H^2\left(\hat{s}_m, \hat{s}_{\hat{m}}\right) &\leq \left(1+\upsilon^{-1}\right) H^2\left(\hat{s}_m, s\mathbb{1}_A\right) \\
&\quad + \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{(1-\varepsilon) H^2\left(s\mathbb{1}_A, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}(m')\right\} + 2\mathrm{pen}(m) + \frac{1}{n} \\
&\leq \left(1+\upsilon^{-1}\right) H^2\left(\hat{s}_m, s\mathbb{1}_A\right) + \left[(1+\varepsilon) H^2\left(s\mathbb{1}_A, \hat{s}_m\right) + \mathrm{pen}(m) + 22\xi\right] + 2\mathrm{pen}(m) + \frac{1}{n} \\
&\leq \left(2+\varepsilon+\upsilon^{-1}\right) H^2\left(\hat{s}_m, s\mathbb{1}_A\right) + 3\mathrm{pen}(m) + 22\xi + \frac{1}{n}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\alpha H^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right) &\leq 2\alpha H^2\left(s\mathbb{1}_A, \hat{s}_m\right) + 2\alpha H^2\left(\hat{s}_m, \hat{s}_{\hat{m}}\right) \\
&\leq 2\left(2+\alpha+\varepsilon+\upsilon^{-1}\right) H^2\left(\hat{s}_m, s\mathbb{1}_A\right) + 6\mathrm{pen}(m) + 44\xi + \frac{2}{n}.
\end{aligned}$$

Finally, we have proved that there exists $C > 0$, such that, with probability larger than $1 - 3e^{-n\xi}$, for all $m \in \mathcal{M}_\ell$,

$$C H^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right) \leq H^2\left(\hat{s}_m, s\mathbb{1}_A\right) + \mathrm{pen}(m) + \xi.$$

This concludes the proof. □

6.3.1. Proof of Lemma 5. For each partition $m \in \mathcal{M}_\ell$, we define the random set $\hat{S}_m$ of functions as follows. A function $\hat{f}$ is said to belong to $\hat{S}_m$ if for each cube $K \in m$, there exists a partition $m_K \in \mathcal{M}_\ell$ such that $\hat{f} = \hat{s}_{m_K}$ on K. In other words,

$$\hat{S}_m = \left\{ \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K,\ \forall K \in m,\ m_K \in \mathcal{M}_\ell \right\}.$$

For all function $\hat{f} \in \hat{S}_m$ and $K \in m$, let $m_K(\hat{f}) \in \mathcal{M}_\ell$ be any partition such that

$$|m_K(\hat{f}) \vee K| = \inf\left\{ |m' \vee K|,\ m' \in \mathcal{M}_\ell,\ \hat{f}\mathbb{1}_K = \hat{s}_{m'}\mathbb{1}_K \right\}.$$

The function $\hat{f} \in \hat{S}_m$ is piecewise constant on the elements of the partition

$$m(\hat{f}) = \bigcup_{K \in m} (m_K(\hat{f}) \vee K).$$

The whole point is that

$$\begin{aligned}
\gamma(m) &= \left\{ \sum_{K \in m} \sup_{m'_K \in \mathcal{M}_\ell} \left[\alpha H^2\big(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'_K} \mathbb{1}_K\big) + T\big(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'_K} \mathbb{1}_K\big) - \mathrm{pen}(m'_K \vee K)\right] \right\} + 2\mathrm{pen}(m) \\
&= \sup_{\hat{f} \in \hat{S}_m} \left\{ \sum_{K \in m} \left[\alpha H^2\big(\hat{s}_m \mathbb{1}_K, \hat{f}\mathbb{1}_K\big) + T\big(\hat{s}_m \mathbb{1}_K, \hat{f}\mathbb{1}_K\big) - \mathrm{pen}(m_K(\hat{f}) \vee K)\right] \right\} + 2\mathrm{pen}(m).
\end{aligned}$$

Since $|m(\hat{f})| = \sum_{K\in m} |m_K(\hat{f}) \vee K|$,

$$\gamma(m) = \sup_{\hat{f}\in\hat{S}_m} \left\{ \alpha H^2\big(\hat{s}_m, \hat{f}\big) + T\big(\hat{s}_m, \hat{f}\big) - \mathrm{pen}(m(\hat{f})) \right\} + 2\mathrm{pen}(m). \tag{15}$$

We can now prove the first part of the lemma. For all partitions $m, m' \in \mathcal{M}_\ell$, the estimator $\hat{s}_{m'}$ belongs to $\hat{S}_m$. Hence,

$$\gamma(m) \geq \sup_{m'\in\mathcal{M}_\ell} \left\{ \alpha H^2\left(\hat{s}_m, \hat{s}_{m'}\right) + T\left(\hat{s}_m, \hat{s}_{m'}\right) - \mathrm{pen}(m(\hat{s}_{m'})) \right\} + 2\mathrm{pen}(m).$$

Since $|m_K(\hat{s}_{m'}) \vee K| \leq |m' \vee K|$, $|m(\hat{s}_{m'})| \leq \sum_{K\in m} |m' \vee K|$ and

$$\sum_{K\in m} |m' \vee K| = \left|\left\{K \cap K',\, (K,K') \in m \times m',\, K \cap K' \neq \emptyset\right\}\right|.$$

Remark that either $K \cap K' = K$ or $K \cap K' = K'$ since K, K' are non-disjoint cubes (see Figure 3.1). Hence, $|m(\hat{s}_{m'})| \leq |m| + |m'|$, and $\gamma_1(m) \leq \gamma(m)$ as wished.

To prove the second part of the lemma, we must define the set S_m appearing in the definition of γ_2.

Definition 1. *Let, for all $K \in \cup_{m\in\mathcal{M}_\ell} m$, $K_0, \ldots, K_l$ be the cubes of $\cup_{m\in\mathcal{M}_\ell} m$ such that $K \subset K_i$ for all $i \in \{0,\ldots,l\}$. For all $i \in \{0,\ldots,l\}$, let I_i and J_i be the subsets of $[0,1]^d$ such that $K_i = I_i \times J_i$. Set*

$$S_K = \bigcup_{i=0}^{l} \left\{ \frac{a}{b\mu(J_i)} \mathbb{1}_K,\, a \in \{0,\ldots,n\},\, b \in \{1,\ldots,n\} \right\}$$

with the convention $a/0 = 0$ whatever $a \in \{0,\ldots,n\}$. We define S_m by

$$S_m = \left\{ \sum_{K\in m} f_K,\, f_K \in S_K \right\}.$$

Let us now make the link between S_m and $\hat{S}_m$. We shall show that a function $\hat{f} \in \hat{S}_m$ belongs to $S_{m(\hat{f})}$. For this purpose, let $K' \in m(\hat{f})$. By definition of $m(\hat{f})$, there exist $K \in m$, $K'' \in m_K(\hat{f})$ such that $K' = K'' \cap K$. We have $\hat{f}\mathbb{1}_K = \hat{s}_{m_K(\hat{f})}\mathbb{1}_K$ and

$$\hat{s}_{m_K(\hat{f})}\mathbb{1}_{K''} = \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K''}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_{K''}(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_{K''}.$$

Consequently,

$$\begin{aligned} \hat{f}\mathbb{1}_{K'} &= \big(\hat{f}\mathbb{1}_K\big)\mathbb{1}_{K''} \\ &= \big(\hat{s}_{m_K(\hat{f})}\mathbb{1}_K\big)\mathbb{1}_{K''} = \big(\hat{s}_{m_K(\hat{f})}\mathbb{1}_{K''}\big)\mathbb{1}_K \\ &= \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K''}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_{K''}(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_{K'}. \end{aligned}$$

This implies that $\hat{f}\mathbb{1}_{K'} \in S_{K'}$ and thus $\hat{f} = \sum_{K' \in m(\hat{f})} \hat{f}\mathbb{1}_{K'}$ belongs to $S_{m(\hat{f})}$.

The inequality $\gamma_2 \geq \gamma$ ensues from (15), $\hat{S}_m \subset \cup_{m' \in \mathcal{M}_\ell} S_{m'}$ and

$$\gamma_2(m) = \sup_{f' \in \bigcup_{m' \in \mathcal{M}_\ell} S_{m'}} \left\{ \alpha H^2\left(\hat{s}_m, f'\right) + T\left(\hat{s}_m, f'\right) - \left(\inf_{\substack{m'' \in \mathcal{M}_\ell \\ S_{m''} \ni f'}} \mathrm{pen}(m'') \right) \right\} + 2\mathrm{pen}(m).$$

□

6.3.2. Proof of Lemma 6. We start with the claim below.

Claim 3. *Let ψ be the bounded function defined on $[0,+\infty)^2$ by*

$$\psi(x,y) = \frac{1}{\sqrt{2}} \frac{\sqrt{y} - \sqrt{x}}{\sqrt{x+y}} \quad \text{for all } x, y \in [0,+\infty)$$

with the convention $0/0 = 0$.

Let, for all $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M)$, with support included in A, $Z(f,f')$ be the random variable defined by

$$Z(f,f') = \frac{1}{n} \sum_{i=0}^{n-1} \left(\psi\left(f(X_i, X_{i+1}), f'(X_i, X_{i+1})\right) - \int_{\mathbb{X}} \psi\left(f(X_i,y), f'(X_i,y)\right) (s\mathbb{1}_A)(X_i,y)\, \mathrm{d}\mu(y) \right).$$

Then,

$$\left(1 - \frac{1}{\sqrt{2}}\right) H^2\left(s\mathbb{1}_A, f'\right) + T\left(f, f'\right) \leq \left(1 + \frac{1}{\sqrt{2}}\right) H^2\left(s\mathbb{1}_A, f\right) + Z\left(f, f'\right) \tag{16}$$

and

$$\frac{1}{n} \sum_{i=0}^{n-1} \int_{\mathbb{X}} \psi^2\left(f(X_i,y), f'(X_i,y)\right) (s\mathbb{1}_A)(X_i,y)\, \mathrm{d}\mu(y) \leq 3\left(H^2(s\mathbb{1}_A, f) + H^2(s\mathbb{1}_A, f')\right). \tag{17}$$

Proof. For all $i \in \{0, \ldots, n-1\}$, let

$$\begin{aligned}
T_i(f,f') &= \frac{1}{2\sqrt{2}} \int_{\mathbb{X}} \sqrt{f(X_i,y) + f'(X_i,y)} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)} \right) \mathrm{d}\mu(y) \\
&\quad + \frac{1}{\sqrt{2}} \frac{\sqrt{f'(X_i,X_{i+1})} - \sqrt{f(X_i,X_{i+1})}}{\sqrt{f(X_i,X_{i+1}) + f'(X_i,X_{i+1})}} + \frac{1}{2} \int_{\mathbb{X}} \left(f(X_i,y) - f'(X_i,y) \right) \mathrm{d}\mu(y) \\
Z_i(f,f') &= \psi\left(f(X_i,X_{i+1}), f'(X_i,X_{i+1})\right) - \mathbb{E}\left[\psi\left(f(X_i,X_{i+1}), f'(X_i,X_{i+1})\right) \mid X_i\right] \\
H_i^2(f,f') &= \frac{1}{2} \int_{\mathbb{X}} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)} \right)^2 \mathrm{d}\mu(y) \\
\rho_i(f,f') &= \int_{\mathbb{X}} \sqrt{f'(X_i,y) f(X_i,y)}\, \mathrm{d}\mu(y).
\end{aligned}$$

Since

$$T(f,f')=\frac{1}{n}\sum_{i=0}^{n-1}T_i(f,f')\ ,\quad Z(f,f')=\frac{1}{n}\sum_{i=0}^{n-1}Z_i(f,f')\text{ and }\quad H^2(f,f')=\frac{1}{n}\sum_{i=0}^{n-1}H_i^2(f,f')$$

it is sufficient to prove that for all $i\in\{0,\dots,n-1\}$,

$$\left(1-\frac{1}{\sqrt{2}}\right)H_i^2\left(s\mathbb{1}_A,f'\right)+T_i\left(f,f'\right)\leq\left(1+\frac{1}{\sqrt{2}}\right)H_i^2\left(s\mathbb{1}_A,f\right)+Z_i\left(f,f'\right)\tag{18}$$

and

$$\int_{\mathbb{X}}\psi^2\left(f(X_i,y),f'(X_i,y)\right)(s\mathbb{1}_A)(X_i,y)\,\mathrm{d}\mu(y)\leq 3\left(H_i^2(s\mathbb{1}_A,f)+H_i^2(s\mathbb{1}_A,f')\right).\tag{19}$$

These inequalities are established by using similar arguments than those developed in the proofs of Propositions 2 and 3 of Baraud (2011). Let us make them explicit. We set

$$\rho_i(\zeta,f,f')=\frac{1}{2}\left[\rho_i\left(f',\frac{f+f'}{2}\right)+\int_{\mathbb{X}}\sqrt{\frac{2f'(X_i,y)}{f(X_i,y)+f'(X_i,y)}}\,\mathrm{d}\zeta(y)\right]\quad\text{for all measure }\zeta$$

where the convention $0/0$ is in use. Let ζ_i be the random measure defined by $\mathrm{d}\zeta_i/\,\mathrm{d}\mu=(s\mathbb{1}_A)(X_i,\cdot)$. Then,

$$\mathbb{E}\left[T_i(f,f')\mid X_i\right]=\left[\rho_i(\zeta_i,f,f')-\frac{1}{2}\int_{\mathbb{X}}f'(X_i,y)\,\mathrm{d}\mu(y)\right]-\left[\rho_i(\zeta_i,f',f)-\frac{1}{2}\int_{\mathbb{X}}f(X_i,y)\,\mathrm{d}\mu(y)\right].$$

We deduce from relation (6) of Baraud (2011),

$$\begin{aligned}0&\leq\rho_i(\zeta_i,f,f')-\rho_i(s\mathbb{1}_A,f')\leq\frac{1}{\sqrt{2}}\left[H_i^2(s\mathbb{1}_A,f)+H_i^2(s\mathbb{1}_A,f')\right]\\0&\leq\rho_i(\zeta_i,f',f)-\rho_i(s\mathbb{1}_A,f)\leq\frac{1}{\sqrt{2}}\left[H_i^2(s\mathbb{1}_A,f)+H_i^2(s\mathbb{1}_A,f')\right].\end{aligned}$$

Now,

$$\begin{aligned}H_i^2(f',s\mathbb{1}_A)-H_i^2(f,s\mathbb{1}_A)&=\left[\rho_i(s\mathbb{1}_A,f)-\frac{1}{2}\int_{\mathbb{X}}f(X_i,y)\,\mathrm{d}\mu(y)\right]\\&\qquad-\left[\rho_i(s\mathbb{1}_A,f')-\frac{1}{2}\int_{\mathbb{X}}f'(X_i,y)\,\mathrm{d}\mu(y)\right]\\&=-\mathbb{E}\left[T_i(f,f')\mid X_i\right]+\left[\rho_i(\zeta_i,f,f')-\rho_i(s\mathbb{1}_A,f')\right]\\&\qquad-\left[\rho_i(\zeta_i,f',f)-\rho_i(s\mathbb{1}_A,f)\right]\\&\leq-\mathbb{E}\left[T_i(f,f')\mid X_i\right]+\frac{1}{\sqrt{2}}\left[H_i^2(s\mathbb{1}_A,f)+H_i^2(s\mathbb{1}_A,f')\right].\end{aligned}$$

Inequality (18) ensues from the relation $Z_i(f,f')=T_i(f,f')-\mathbb{E}\left[T_i(f,f')\mid X_i\right]$. Let us now

prove (19).

$$
\begin{aligned}
&2\int_{\mathbb{X}} \psi^2\left(f(X_i,y), f'(X_i,y)\right) s\mathbb{1}_A(X_i,y)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{X}} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right)^2 \frac{s\mathbb{1}_A(X_i,y)}{f(X_i,y)+f'(X_i,y)}\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{X}} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right)^2 \left(\sqrt{\frac{s\mathbb{1}_A(X_i,y)}{f(X_i,y)+f'(X_i,y)}} - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right)^2 \mathrm{d}\mu(y)\\
&\le 2\int_{\mathbb{X}} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right)^2 \left(\sqrt{\frac{s\mathbb{1}_A(X_i,y)}{f(X_i,y)+f'(X_i,y)}} - \frac{1}{\sqrt{2}}\right)^2 \mathrm{d}\mu(y)\\
&\qquad + \int_{\mathbb{X}} \left(\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}\right)^2 \mathrm{d}\mu(y)\\
&\le 2\int_{\mathbb{X}} \left(\frac{\sqrt{f'(X_i,y)} - \sqrt{f(X_i,y)}}{\sqrt{f(X_i,y)+f'(X_i,y)}}\right)^2 \left(\sqrt{s\mathbb{1}_A(X_i,y)} - \sqrt{\frac{f(X_i,y)+f'(X_i,y)}{2}}\right)^2 \mathrm{d}\mu(y)\\
&\qquad + 2H_i^2(f,f').
\end{aligned}
$$

The first term of the right-hand side of this inequality can be upper bounded by $4H_i^2\left(s\mathbb{1}_A, \frac{f+f'}{2}\right)$. Since the map $t \mapsto \sqrt{t}$ is concave, $2H_i^2\left(s\mathbb{1}_A, \frac{f+f'}{2}\right) \le H_i^2(s\mathbb{1}_A, f) + H_i^2(s\mathbb{1}_A, f')$. The second term is bounded from above by $H_i^2(f,f') \le 2(H_i^2(s\mathbb{1}_A, f) + H_i^2(s\mathbb{1}_A, f'))$. Finally,

$$
\int_{\mathbb{X}} \psi^2\left(f(X_i,y), f'(X_i,y)\right) s\mathbb{1}_A(X_i,y)\,\mathrm{d}\mu(y) \le 3\left[H_i^2(s\mathbb{1}_A, f) + H_i^2(s\mathbb{1}_A, f')\right]
$$

as wished. □

We shall prove (14) by applying the following concentration inequality to the random variable $Z(f,f')$.

Claim 4. *For all $i \le n-1$, let $\mathcal{F}_i$ be the σ-field generated by the random variables X_j for $j \in \{0,\dots,i\}$. Let $f_0,\dots,f_{n-1} \in \mathbb{L}^1(\mathbb{X}^2, M)$ such that there exists $b \in \mathbb{R}$ with $\sup_{x\in\mathbb{X}^2} |f_i(x)| \le b$ for all $i \in \{0,\dots,n-1\}$. Set*

$$
S_n = \sum_{i=0}^{n-1} \left(f_i(X_i, X_{i+1}) - \mathbb{E}\left[f_i(X_i, X_{i+1}) \mid \mathcal{F}_i\right]\right)
$$

and

$$
V_n = \sum_{i=0}^{n-1} \mathbb{E}\left[f_i^2(X_i, X_{i+1}) \mid \mathcal{F}_i\right].
$$

Then, for all $\beta > b$ and $x > 0$

$$
\mathbb{P}\left[S_n \ge \frac{V_n}{2(\beta-b)} + \beta x\right] \le e^{-x}.
$$

Proof. By setting $a^{-1} = 2(\beta - b)$,

$$\begin{aligned}\log \mathbb{P}\left[S_n \geq aV_n + \beta x\right] &\leq -x + \log \mathbb{E}\left[\exp\left(\beta^{-1}S_n - a\beta^{-1}V_n\right)\right] \\ &\leq -x + \log \mathbb{E}\left[\exp\left(\beta^{-1}S_{n-1} - a\beta^{-1}V_n\right)\mathbb{E}\left[\exp\left(\beta^{-1}(S_n - S_{n-1})\right) \mid \mathcal{F}_{n-1}\right]\right].\end{aligned}$$

By using Bernstein inequality (equation (2.21) of Massart (2003)),

$$\mathbb{E}\left[\exp\left(\beta^{-1}(S_n - S_{n-1})\right) \mid \mathcal{F}_{n-1}\right] \leq \exp\left(\frac{\beta^{-2}(V_n - V_{n-1})}{2(1-\beta^{-1}b)}\right)$$

and thus

$$\log \mathbb{P}\left[S_n \geq aV_n + \beta x\right] \leq -x + \log \mathbb{E}\left[\exp\left(\beta^{-1}S_{n-1} - a\beta^{-1}V_{n-1}\right)\right].$$

The result follows by induction. □

Proof of Lemma 6. Set $z = (1 - 1/\sqrt{2})/4$, $\beta = (3/z + \sqrt{2})/2$ and for all $\xi > 0$,

$$\Omega_\xi = \left\{ \sup_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} \frac{Z(f,f')}{z\left(H^2(f,s\mathbb{1}_A) + H^2(f',s\mathbb{1}_A)\right) + \mathrm{pen}(m) + \mathrm{pen}(m') + \beta\xi} < 1 \right\}.$$

On Ω_ξ, for all $m, m' \in \mathcal{M}_\ell$, $(f, f') \in S_m \times S_{m'}$,

$$Z(f,f') \leq z\left(H^2(f,s\mathbb{1}_A) + H^2(f',s\mathbb{1}_A)\right) + \mathrm{pen}(m) + \mathrm{pen}(m') + \beta\xi$$

and (14) derives from (16) (with $\varepsilon = 1/\sqrt{2} + z$).

It remains to prove that $\mathbb{P}\left(\Omega_\xi^c\right) \leq 3e^{-n\xi}$. We have

$$\mathbb{P}\left(\Omega_\xi^c\right) \leq \sum_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} \mathbb{P}\left[Z(f,f') \geq z\left[H^2(s\mathbb{1}_A,f) + H^2(s\mathbb{1}_A,f')\right] + \mathrm{pen}(m) + \mathrm{pen}(m') + \beta\xi\right].$$

We apply the concentration inequality given by Claim 4 with $f_i = \psi(f,f')$, $b = 1/\sqrt{2}$, $S_n = nZ(f,f')$ and by using relation (17),

$$V_n = \sum_{i=0}^{n-1}\int_{\mathbb{X}} \psi^2\left(f(X_i,y), f'(X_i,y)\right)(s\mathbb{1}_A)(X_i,y)\,\mathrm{d}\mu(y) \leq 3n\left(H^2(s\mathbb{1}_A,f) + H^2(s\mathbb{1}_A,f')\right).$$

We obtain for all $x > 0$,

$$\mathbb{P}\left[Z(f,f') \geq \frac{3}{\sqrt{2}(\beta\sqrt{2}-1)}\left[H^2(s\mathbb{1}_A,f) + H^2(s\mathbb{1}_A,f')\right] + \beta\frac{x}{n}\right] \leq e^{-x}.$$

Note that $z = 3/(\sqrt{2}(\beta\sqrt{2}-1))$. By using the above inequality with

$$\beta\frac{x}{n} = \mathrm{pen}(m) + \mathrm{pen}(m') + \beta\xi$$

we deduce that

$$\mathbb{P}\left(\Omega_\xi^c\right) \leq \sum_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} e^{-n\left(\beta^{-1}\mathrm{pen}(m)+\beta^{-1}\mathrm{pen}(m')+\xi\right)}.$$

Now, the set S_K defined in Definition 1 page 88, satisfies $|S_K| \leq (\ell+1)n(n+1)$ which implies that $|S_m| \leq ((\ell+1)n(n+1))^{|m|}$. By using $\ell \leq n$, $\log|S_m| \leq 3|m|\log(n+1)$. Since L is large enough ($L \geq 90$), $\beta^{-1}\mathrm{pen}(m) \geq (|m| + \log|S_m|)/n$ for all $m \in \mathcal{M}_\ell$. Consequently,

$$\begin{aligned}\mathbb{P}\left(\Omega_\xi^c\right) &\leq \sum_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} e^{-(|m|+\log|S_m|+|m'|+\log|S_{m'}|+n\xi)}\\ &\leq \left(\sum_{m\in\mathcal{M}_\ell} e^{-|m|}\right)^2 e^{-n\xi}.\end{aligned}$$

The conclusion follows from the inequality $\sum_{m\in\mathcal{M}_\ell} e^{-|m|} \leq \sqrt{3}$ (see Section 3.2.4 of Baraud and Birgé (2009)). □

6.3.3. Proof of Lemma 4. The proof of this lemma is based on the two following remarks.

1. The Hellinger distance $H^2(f,f')$ and the test $T(f,f')$ are respectively upper bounded by 1 and 2 when f, f' are such that

$$\frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{R}^d} f(X_i,y)\,\mathrm{d}\mu(y) \leq 1 \quad \text{and} \quad \frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{R}^d} f'(X_i,y)\,\mathrm{d}\mu(y) \leq 1.$$

2. The cardinality of a partition $m \in \mathcal{M}_\ell \setminus \mathcal{M}_n$ is lower bounded by $|m| \geq n+1$ when $\ell \geq n+1$.

More precisely, the proof follows from the two claims below.

Claim 5. *Let for each $m_1, m_2 \in \mathcal{M}_\infty$ and $K \in m_1$,*

$$\gamma_K(m_1,m_2) = \alpha H^2\left(\hat{s}_{m_1}\mathbb{1}_K, \hat{s}_{m_2}\mathbb{1}_K\right) + T\left(\hat{s}_{m_1}\mathbb{1}_K, \hat{s}_{m_2}\mathbb{1}_K\right) - \mathrm{pen}(m_2 \vee K).$$

Then, for all $\ell \in \mathbb{N}^\star$, $\ell \geq n+1$, $m_1 \in \mathcal{M}_\infty$, $K \in m_1$,

$$\sup_{m_2\in\mathcal{M}_\ell} \gamma_K(m_1,m_2) = \sup_{m_2\in\mathcal{M}_n} \gamma_K(m_1,m_2)$$

and thus

$$\sup_{m_2\in\mathcal{M}_\infty} \gamma_K(m_1,m_2) = \sup_{m_2\in\mathcal{M}_n} \gamma_K(m_1,m_2).$$

Proof. Let $m_2^\star \in \mathcal{M}_\ell$ such that $\gamma_K(m_1,m_2^\star) = \sup_{m_2\in\mathcal{M}_\ell}\gamma_K(m_1,m_2)$. In Section 2, we have defined the collection $\mathcal{M}_\ell$ of partitions of $[0,1]^{2d}$. Likewise, by using the algorithm of DeVore and Yu

(1990), we define the collection $\mathcal{M}_\ell(K)$ of partitions of K. Note that $m_2^\star \vee K$ belongs to $\mathcal{M}_\ell(K)$. Since $H^2\left(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K\right) \leq 1$ and $\left|T\left(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K\right)\right| \leq 2$, we have

$$\gamma_K(m_1, m_2^\star) \leq 3 - L\frac{|m_2^\star \vee K| \log n}{n}.$$

Remark that

$$\gamma_K(m_1, m_2^\star) \geq \gamma_K\big(m_1, \{[0,1]^{2d}\}\big) \geq -2 - L\frac{\log n}{n}$$

which leads to

$$|m_2^\star \vee K| \leq 1 + \frac{5n}{L \log n} \leq n.$$

This implies that $m_2^\star \vee K$ belongs to $\mathcal{M}_n(K)$. There exists $m_2^\bullet \in \mathcal{M}_n$ such that $m_2^\bullet \vee K = m_2^\star \vee K$ and hence $\gamma_K(m_1, m_2^\bullet) = \gamma_K(m_1, m_2^\star)$ which concludes the proof. □

Claim 6. *Set for all $m \in \mathcal{M}_\infty$ and $K \in m$,*

$$\gamma_K(m) = \sup_{m_2 \in \mathcal{M}_\ell} \gamma_K(m, m_2).$$

Then, $\gamma(m) = 2\mathrm{pen}(m) + \sum_{K \in m} \gamma_K(m)$ and for all $\ell \in \mathbb{N}^\star$, $\ell \geq n+1$,

$$\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m)$$

and thus

$$\inf_{m \in \mathcal{M}_\infty} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m).$$

Proof. Let $m^\star \in \mathcal{M}_\ell$ such that $\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \gamma(m^\star)$. By Lemma 5,

$$\begin{aligned}
\gamma(m^\star) &\geq \sup_{m' \in \mathcal{M}_\ell} \left\{\alpha H^2\left(\hat{s}_{m^\star}, \hat{s}_{m'}\right) + T\left(\hat{s}_{m^\star}, \hat{s}_{m'}\right) - \mathrm{pen}(m')\right\} + L\frac{|m^\star| \log n}{n} \\
&\geq \left(-2 - L\frac{\log n}{n}\right) + L\frac{|m^\star| \log n}{n} \\
&\geq -2 + L\frac{(|m^\star| - 1) \log n}{n}.
\end{aligned}$$

Now,

$$\gamma(m^\star) \leq \gamma(\{[0,1]^{2d}\}) \leq 3 + 2L\frac{\log n}{n}$$

which implies that

$$|m^\star| \leq 3 + \frac{5n}{L \log n} \leq n$$

and thus $m^\star \in \mathcal{M}_n$. □

6.4. Proof of Theorem 3. Consider the regular partition m_{ref} of $[0,1]^{2d}$ into cubes with side length $2^{-\ell}$, that is

$$m_{ref} = \left\{ K_{\ell,\mathbf{l}}, \mathbf{l} = (k,\dots,k),\ k \in \{1,\dots,2^\ell\} \right\}$$

where $K_{\ell,\mathbf{l}}$ is defined in Section 2.2. For all partition $m \in \mathcal{M}_\ell$, $V_m \subset V_{m_{ref}}$. Set

$$\Omega_{eq} = \left[\forall g_1, g_2 \in V_{m_{ref}},\ h^2(g_1,g_2) \le 11 H^2(g_1,g_2) \right]$$

and define $\bar{s}_m$ an element of V_m such that $h^2(s\mathbb{1}_A, \bar{s}_m) = h^2(s\mathbb{1}_A, V_m)$.

For all $m \in \mathcal{M}_\ell$,

$$\begin{aligned}
\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] &\le \mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}}\right] + \mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right] \\
&\le 2\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \bar{s}_m\right)\mathbb{1}_{\Omega_{eq}}\right] + 2\mathbb{E}\left[h^2\left(\bar{s}_m, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}}\right] + \mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right] \\
&\le 2\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \bar{s}_m\right)\mathbb{1}_{\Omega_{eq}}\right] + 22\mathbb{E}\left[H^2\left(\bar{s}_m, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}}\right] + \mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right] \\
&\le 2\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \bar{s}_m\right)\mathbb{1}_{\Omega_{eq}}\right] + 44\mathbb{E}\left[H^2\left(s\mathbb{1}_A, \bar{s}_m\right)\mathbb{1}_{\Omega_{eq}}\right] + 44\mathbb{E}\left[H^2\left(\hat{s}_{\hat{m}}, s\mathbb{1}_A\right)\mathbb{1}_{\Omega_{eq}}\right] \\
&\quad + \mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right].
\end{aligned}$$

Now, $h^2\left(s\mathbb{1}_A, \bar{s}_m\right) = \mathbb{E}[H^2\left(s\mathbb{1}_A, \bar{s}_m\right)] = h^2(s\mathbb{1}_A, V_m)$ and

$$\begin{aligned}
\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right] &\le 2\mathbb{E}\left[\left(h^2(s,0) + h^2(\hat{s}_{\hat{m}},0)\right)\mathbb{1}_{\Omega_{eq}^c}\right] \\
&\le \mathbb{E}\left[\left(1 + 2\sup_{m\in\mathcal{M}_\ell} h^2(\hat{s}_m,0)\right)\mathbb{1}_{\Omega_{eq}^c}\right].
\end{aligned}$$

Let for all $K \in m$, I_K and J_K be the subsets of $[0,1]^d$ such that $K = I_K \times J_K$. Then,

$$2h^2(\hat{s}_m,0) = \sum_{K\in m} \frac{\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)\mathbb{1}_{J_K}(X_{i+1})}{\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)} \int_{I_K} \varphi(x)\,\mathrm{d}x \le |m|.$$

Since $m \subset m_{ref}$, $|m| \le |m_{ref}| = 4^{\ell d}$ and thus

$$\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_{eq}^c}\right] \le (1 + 4^{\ell d})\mathbb{P}\left(\Omega_{eq}^c\right) \le 2 \times 4^{\ell d}\mathbb{P}\left(\Omega_{eq}^c\right).$$

We have proved that there exists an universal constant $C' > 0$ such that

$$C'\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] \le \inf_{m\in\mathcal{M}_\ell}\left\{h^2\left(s\mathbb{1}_A, V_m\right) + \mathrm{pen}(m)\right\} + 4^{\ell d}\mathbb{P}\left(\Omega_{eq}^c\right).$$

We now bound from above the term $\mathbb{P}\left(\Omega_{eq}^c\right)$. We denote by $\mathbf{I}_{ref}$ the regular partition of $[0,1]^d$ into cubes with side length $2^{-\ell}$. Remark that

$$\begin{aligned}
\mathbb{P}\left(\Omega_{eq}^c\right) &\le \mathbb{P}\left[\exists I \in \mathbf{I}_{ref},\ \mathbb{P}\left(X_1 \in I\right) \ge \frac{11}{n}\sum_{i=0}^{n-1} \mathbb{1}_I(X_i)\right] \\
&\le 2^{\ell d}\sup_{I\in\mathbf{I}_{ref}} \mathbb{P}\left[\frac{1}{n}\sum_{i=0}^{n-1}\left(\mathbb{1}_I(X_i) - \mathbb{P}\left(X_i \in I\right)\right) \le -\frac{10}{11}\mathbb{P}\left(X_1 \in I\right)\right].
\end{aligned}$$

We use the following Bennett-type inequality for β-mixing random variables (with $f = -\mathbb{1}_I$, $v = \mathbb{P}\left(X_1 \in I\right)$, $c = 0$, $\xi = 10/11\mathbb{P}\left(X_1 \in I\right)$).

Proposition 8. *Let $(X_i)_{i\geq 0}$ be a stationary Markov chain with values in $\mathbb{R}^d$, and let f be a real-valued function on $\mathbb{R}^d$ upper bounded by $c \geq 0$ such that $v = \mathbb{E}\left[f(X_1)^2\right] < \infty$.*

Then, for all $q \in \{1, \ldots, n\}$ and $\xi > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1}\left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right) > \xi\right) \leq 2\exp\left(-\frac{n\xi^2}{8q\left(v + c\xi/6\right)}\right) + 3n\beta_q/q.$$

We then have for all $I \in \mathbf{I}_{ref}$,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=0}^{n-1}\left(\mathbb{1}_I(X_i) - \mathbb{P}\left(X_i \in I\right)\right) \leq -\frac{10}{11}\mathbb{P}\left(X_1 \in I\right)\right] &\leq 3\inf_{1\leq q\leq n}\left\{\exp\left(-\frac{25n\mathbb{P}\left(X_1 \in I\right)}{242q}\right) + n\beta_q/q\right\} \\ &\leq 3\inf_{1\leq q\leq n}\left\{\exp\left(-\frac{n\kappa_0}{10q2^{\ell d}}\right) + n\beta_q/q\right\}\end{aligned}$$

which concludes the proof. □

Proof of Proposition 8. Let l be the smallest integer larger than $n/(2q)$. We derive from Berbee's lemma and more precisely from Viennet (1997) (page 484) that there exist $X_0^\star, \ldots, X_{2lq-1}^\star$ such that

- For $j = 1, \ldots, l$, the random vectors
$$\mathbf{X}_{j,1} = (X_{2(j-1)q}, \ldots, X_{2(j-1)q+q-1}) \quad \text{and} \quad \mathbf{X}_{j,1}^\star = (X_{2(j-1)q}^\star, \ldots, X_{2(j-1)q+q-1}^\star)$$
have the same distribution, and so have the random vectors
$$\mathbf{X}_{j,2} = (X_{2(j-1)q+q}, \ldots, X_{2jq-1}) \quad \text{and} \quad \mathbf{X}_{j,2}^\star = (X_{2(j-1)q+q}^\star, \ldots, X_{2jq-1}^\star).$$
- The random vectors $\mathbf{X}_{1,1}^\star, \ldots, \mathbf{X}_{l,1}^\star$ are independent. The random vectors $\mathbf{X}_{1,2}^\star, \ldots, \mathbf{X}_{l,2}^\star$ are also independent.
- The event
$$\Omega^\star = \bigcap_{1\leq j\leq l}\left(\left[\mathbf{X}_{j,1} \neq \mathbf{X}_{j,1}^\star\right] \cap \left[\mathbf{X}_{j,2} \neq \mathbf{X}_{j,2}^\star\right]\right)$$
satisfies $\mathbb{P}\left[(\Omega^\star)^c\right] \leq 2l\beta_q$.

We set $g_i(x) = f(x)$ if $i \leq n-1$ and $g_i(x) = 0$ otherwise. For $j \in \{1, \ldots, l\}$, we set

$$g'_{j,1}(x_0, \ldots, x_{q-1}) = \sum_{i=0}^{q-1} g_{2(j-1)q+i}(x_i) \quad \text{and} \quad g'_{j,2}(x_0, \ldots, x_{q-1}) = \sum_{i=0}^{q-1} g_{2(j-1)q+q+i}(x_i).$$

Then,

$$\begin{aligned}\mathbb{P}\left[\left(\frac{1}{n}\sum_{i=0}^{n-1}\left(g_i(X_i) - \mathbb{E}\left[g_i(X_i)\right]\right) > \xi\right) \cap \Omega^\star\right] &\leq \mathbb{P}\left(\sum_{j=1}^{l}\left(g'_{j,1}(\mathbf{X}_{j,1}^\star) - \mathbb{E}\left[g'_{j,1}(\mathbf{X}_{j,1}^\star)\right]\right) > n\xi/2\right) \\ &\quad + \mathbb{P}\left(\sum_{j=1}^{l}\left(g'_{j,2}(\mathbf{X}_{j,2}^\star) - \mathbb{E}\left[g'_{j,2}(\mathbf{X}_{j,2}^\star)\right]\right) > n\xi/2\right) \\ &\leq 2\exp\left(-\frac{n^2\xi^2}{8q\left(nv + cn\xi/6\right)}\right)\end{aligned}$$

by using Proposition 2.8 and inequality (2.16) of Massart (2003) (in the paper of Massart (2003), the Bennett inequality holds for $b = 0$ when (2.15) is replaced by (2.16)). □

6.5. Proof of Corollary 2. The corollary ensues from the claim below and Theorem 2 of Baraud and Birgé (2009).

Claim 7. *Under Assumption 2, for all $\ell \in \mathbb{N}^\star$ such that $2^{\ell d} \geq n$,*

$$\inf_{m\in\mathcal{M}_\ell}\left\{d_2^2\left(\sqrt{s}|_A, V_m\right) + \frac{|m|\log n}{n}\right\} \leq 4\inf_{m\in\mathcal{M}_\infty}\left\{d_2^2\left(\sqrt{s}|_A, V_m\right) + \frac{|m|\log n}{n}\right\}.$$

Proof. For all partition $m \in \mathcal{M}_\infty$ and cube $K \in m$, we denote by I_K and J_K the cubes of $[0,1]^d$ such that $K = I_K \times J_K$ and set

$$\bar{s}_m = \sum_{K\in m}\frac{\int_K s(x,y)\,\mathrm{d}x\,\mathrm{d}y}{\mu\otimes\mu(K)}\mathbb{1}_K.$$

In this chapter, d_2 stands for the standard euclidean distance of $\mathbb{L}^2([0,1]^{2d}, \mu\otimes\mu)$. In this proof, we make a slight abuse of notations by denoting by d_2 the standard euclidean distance of $\mathbb{L}^2(\mathbb{R}^{2d}, \mu\otimes\mu)$.

Let $m^\star$ be a partition of $\mathcal{M}_\infty$ such that

$$2\inf_{m\in\mathcal{M}_\infty}\left\{d_2^2\left(\sqrt{s}\mathbb{1}_A, V_m\right) + \frac{|m|\log n}{n}\right\} \geq d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\star}\right) + \frac{|m^\star|\log n}{n}.$$

Let $\mathcal{C}$ be the collection $\mathcal{C} = \{K \in m^\star,\ \mu(I_K) \geq 2^{-\ell d}\}$ and let $m^\bullet$ be a partition of $\mathcal{M}_\ell$ containing $\mathcal{C}$ such that

$$|m^\bullet| = \inf\{|m|,\ m \in \mathcal{M}_\ell \text{ such that } m \ni \mathcal{C}\}.$$

Let $A^\bullet$ be the set defined by $A^\bullet = \cup_{K\in\mathcal{C}} K$ and $V_{m^\bullet}^\bullet = \{f\mathbb{1}_{A^\bullet},\ f \in V_{m^\bullet}\}$.

We have,

$$d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}\right) \leq d_2^2\left(\sqrt{s}\mathbb{1}_{A^\bullet}, V_{m^\bullet}^\bullet\right) + d_2^2\left(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0\right)$$

and

$$d_2^2\left(\sqrt{s}\mathbb{1}_{A^\bullet}, V_{m^\bullet}^\bullet\right) \leq d_2^2\left(\sqrt{s}\mathbb{1}_{A^\bullet}, \sqrt{\bar{s}_{m^\bullet}}\mathbb{1}_{A^\bullet}\right) \leq d_2^2\left(\sqrt{s}\mathbb{1}_A, \sqrt{\bar{s}_{m^\star}}\right).$$

By using Lemma 2 of Baraud and Birgé (2009), $d_2^2\left(\sqrt{s}\mathbb{1}_A, \sqrt{\bar{s}_{m^\star}}\right) \leq 2d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\star}\right)$ which shows that

$$d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}\right) \leq 2d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\star}\right) + d_2^2\left(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0\right).$$

Now,

$$\begin{aligned} d_2^2\left(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0\right) &\leq \sum_{K\in m^\star\setminus\mathcal{C}}\int_{I_K}\left(\int_{\mathbb{R}^d} s(x,y)\,\mathrm{d}y\right)\mathrm{d}x \\ &\leq \sum_{K\in m^\star\setminus\mathcal{C}}\mu(I_K) \leq 2^{-\ell d}|m^\star|. \end{aligned}$$

Since $|m^\bullet| \le |m^\star|$ and $2^{-\ell d} \le n^{-1}$, we have

$$d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}\right) + \frac{|m^\bullet|\log n}{n} \le 2d_2^2\left(\sqrt{s}\mathbb{1}_A, V_{m^\star}\right) + \frac{(1+\log n)|m^\star|}{n}$$

which proves the claim. □

6.6. Rates of convergences for h. We prove the result only for geometrically β-mixing chains (the proof for arithmetically β-mixing chains being similar). We use the claim below whose proof is the same than the one of Claim 7.

Claim 8. *Under Assumption 2, for all $\ell \in \mathbb{N}^\star$ such that $2^{\ell d} \ge n/\log^3 n$,*

$$\inf_{m\in\mathcal{M}_\ell}\left\{h^2\left(s\mathbb{1}_A, V_m\right) + \frac{|m|\log n}{n}\right\} \le 4\inf_{m\in\mathcal{M}_\infty}\left\{h^2\left(s\mathbb{1}_A, V_m\right) + \frac{|m|\log^3 n}{n}\right\}.$$

By using this claim and Theorem 1 of Akakpo (2012),

$$C\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] \le \left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}}\left(\frac{\log^3 n}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log^3 n}{n} + \frac{R_n(\ell)}{n} \tag{20}$$

and by using Theorem 2 of Akakpo (2012),

$$C\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] \le \left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}}\left(\frac{\log n}{n} + 2^{-2\ell d\theta}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}$$

where $C>0$ depends only on κ,σ,d,p and where

$$\theta = \frac{d+\sigma}{\sigma}\left(\frac{\sigma}{d} - 2\left(\frac{1}{p} - \frac{1}{2}\right)_+\right).$$

If $\sigma > \sigma_1(p,d)$ then $\theta > 1/2$. There exits thus n_0 (depending only on θ), such that if $n \ge n_0$, $2^{-2\ell\theta d} \le \log n/n$, and hence

$$C'\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] \le \left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}}\left(\frac{\log n}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}.$$

If $n \le n_0$, we deduce from (20),

$$\begin{aligned} C\mathbb{E}\left[h^2\left(s\mathbb{1}_A, \hat{s}_{\hat{m}}\right)\right] &\le \left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}}\left(\frac{\log^3 n_0}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log^3 n_0}{n} + \frac{R_n(\ell)}{n} \\ &\le C''\left[\left|\sqrt{s}\right|_A\Big|_{p,\sigma}^{\frac{2d}{d+\sigma}}\left(\frac{\log n}{n}\right)^{\frac{\sigma}{\sigma+d}} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}\right] \end{aligned}$$

where C'' depends only on σ, d, p. The conclusion ensues from the fact that $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_1. □

6.7. Proof of Proposition 4. We shall use the following lemma whose proof is similar to the one of Lemma 6.

Lemma 7. *Set $\varepsilon = (2+3\sqrt{2})/8$. Under assumptions of Proposition 4, there exists an universal constant $L_0 > 0$ such that for all $L \geq L_0$ and $\xi > 0$,*

$$\forall f, f' \in S, \quad (1-\varepsilon)\, H^2\left(s\mathbb{1}_A, f'\right) + T\left(f, f'\right) \leq (1+\varepsilon)\, H^2\left(s\mathbb{1}_A, f\right) + L\frac{\Delta_S(f) + \Delta_S(f')}{n} + 22\xi$$

with probability larger than $1 - e^{-n\xi}$.

Proof of Proposition 4. By using the above lemma, with probability larger than $1-e^{-n\xi}$, for all $f \in S$,

$$\sup_{f' \in S} \left\{ (1-\varepsilon)\, H^2\left(s\mathbb{1}_A, f'\right) + T\left(f, f'\right) - L\frac{\Delta_S(f')}{n} \right\} \leq (1+\varepsilon)\, H^2\left(s\mathbb{1}_A, f\right) + L\frac{\Delta_S(f)}{n} + 22\xi.$$

Thus, if $T(f,\hat{f}) + L\frac{\Delta_S(f)}{n} - L\frac{\Delta_S(\hat{f})}{n} \geq 0$,

$$\begin{aligned} \alpha H^2(s\mathbb{1}_A, \hat{f}) &\leq (1-\varepsilon) H^2(s\mathbb{1}_A, \hat{f}) + T(f,\hat{f}) - L\frac{\Delta_S(\hat{f})}{n} + L\frac{\Delta_S(f)}{n} \\ &\leq (1+\varepsilon)\, H^2\left(s\mathbb{1}_A, f\right) + 2L\frac{\Delta_S(f)}{n} + 22\xi. \end{aligned}$$

If $T(f,\hat{f}) + L\frac{\Delta_S(f)}{n} - L\frac{\Delta_S(\hat{f})}{n} < 0$,

$$\begin{aligned} \alpha H^2(f,\hat{f}) &\leq \alpha H^2(\hat{f}, f) + T(\hat{f}, f) - L\frac{\Delta_S(f)}{n} + L\frac{\Delta_S(\hat{f})}{n} \\ &\leq \sup_{f' \in S} \left\{ \alpha H^2(\hat{f}, f') + T(\hat{f}, f') - L\frac{\Delta_S(f')}{n} \right\} + L\frac{\Delta_S(\hat{f})}{n} \\ &\leq \wp(\hat{f}) \\ &\leq \wp(f) + \frac{1}{n} \\ &\leq \sup_{f' \in S} \left\{ \alpha H^2\left(f, f'\right) + T\left(f, f'\right) - L\frac{\Delta_S(f')}{n} \right\} + L\frac{\Delta_S(f)}{n} + \frac{1}{n}. \end{aligned}$$

With $\upsilon = (1-\varepsilon)/\alpha - 1 > 0$,

$$\begin{aligned} \alpha H^2(f,\hat{f}) &\leq \left(1+\upsilon^{-1}\right) H^2\left(f, s\mathbb{1}_A\right) \\ &\qquad + \sup_{f' \in S} \left\{ (1-\varepsilon) H^2\left(s\mathbb{1}_A, f'\right) + T\left(f, f'\right) - L\frac{\Delta_S(f')}{n} \right\} + L\frac{\Delta_S(f)}{n} + \frac{1}{n} \\ &\leq \left(1+\upsilon^{-1}\right) H^2\left(f, s\mathbb{1}_A\right) + \left[(1+\varepsilon)\, H^2\left(s\mathbb{1}_A, f\right) + L\frac{\Delta_S(f)}{n} + 22\xi \right] + L\frac{\Delta_S(f)}{n} + \frac{1}{n} \\ &\leq \left(2+\varepsilon+\upsilon^{-1}\right) H^2\left(f, s\mathbb{1}_A\right) + 2L\frac{\Delta_S(f)}{n} + 22\xi + \frac{1}{n}. \end{aligned}$$

This leads to

$$\begin{aligned}\alpha H^2(s\mathbb{1}_A,\hat{f}) &\leq 2\alpha H^2\left(s\mathbb{1}_A,f\right)+2\alpha H^2(f,\hat{f})\\ &\leq 2\left(2+\alpha+\varepsilon+\upsilon^{-1}\right)H^2\left(f,s\mathbb{1}_A\right)+4L\frac{\Delta_S(f)}{n}+44\xi+\frac{2}{n}.\end{aligned}$$

Finally, we have proved that there exists $C>0$, such that, with probability larger than $1-e^{-n\xi}$, for all $f\in S$,

$$CH^2(s\mathbb{1}_A,\hat{f})\leq H^2\left(f,s\mathbb{1}_A\right)+L\frac{\Delta_S(f)}{n}+\xi.$$

The conclusion follows. □

6.8. Proof of Corollary 4. Throughout this proof, the distance associated to the supremum norm $\|\cdot\|_\infty$ is denoted by d_∞. As defined page 69, d_2 is the usual distance of the space of square integrable functions on $[0,1]^2$ with respect to the Lebesgue measure $\mu\otimes\mu$. We make a slight abuse of notation in this proof since d_2 will also stand for the distance of the space of square integrable functions on $[0,1]$ with respect to the Lebesgue measure μ.

We shall use the following lemma (the first part may be deduced from the work of Akakpo (2012) whereas the second part may be deduced from results in Dahmen et al. (1980)).

Lemma 8. *There exist a collection $\mathbb{W}$ of (finite dimensional) linear spaces and a non-negative map $\Delta_{\mathbb{W}}$ on $\mathbb{W}$ such that $\sum_{W\in\mathbb{W}}e^{-\Delta_{\mathbb{W}}(W)}\leq 1$ and such that for all $p\in(0,+\infty]$, $\beta>(1/p-1/2)_+$ and $f\in\mathscr{B}^\beta(\mathbb{L}^p([0,1]))$, $L>0$, $\tau>0$, $\sigma>0$,*

$$C\inf_{W\in\mathbb{W}}\left\{L^2d_2^{2\sigma}(f,W)+(\dim W+\Delta_{\mathbb{W}}(W))\,\tau\right\}\leq\left(L|f|_{p,\beta}^{\sigma}\right)^{\frac{2}{2\sigma\beta+1}}\tau^{\frac{2\sigma\beta}{2\sigma\beta+1}}+\tau$$

where $C>0$ depends only on p, β. Moreover, for all $\beta>0$, $f\in\mathcal{H}^\beta([0,1])$, $L>0$, $\tau>0$, $\sigma>0$,

$$C'\inf_{W\in\mathbb{W}}\left\{L^2d_\infty^{2\sigma}(f,W)+(\dim W+\Delta_{\mathbb{W}}(W))\,\tau\right\}\leq\left(L|f|_{\infty,\beta}^{\sigma}\right)^{\frac{2}{2\sigma\beta+1}}\tau^{\frac{2\sigma\beta}{2\sigma\beta+1}}+\tau$$

where $C'>0$ depends only on β.

Let us define

$$u(x,y)=\frac{y-g(x)}{1+\|g\|_\infty}\quad\text{and}\quad\Phi(x)=\phi\left((1+\|g\|_\infty)x\right)\quad\text{for all } x,y\in[0,1].$$

Let $\mathbb{W}$ be the family of linear spaces given by the above lemma. Let for all $f\in\cup_{W\in\mathbb{W}}W$, $a\in\mathbb{R}$, $\psi_{a,f}$ be the function defined on $[0,1]^2$ by $\psi_{a,f}(x,y)=a(y-f(x))$. Define, for all $W\in\mathbb{W}$, the linear space

$$T_W=\{\psi_{a,f},\,a\in\mathbb{R},\,f\in W\}.$$

We deduce from the proof of Theorem 2 of Baraud and Birgé (2011) (with $\mathbb{F}=\mathbb{W}$, $l=1$, $\mathbb{T}_1=\{T_W,\,W\in\mathbb{W}\}$, $\gamma(W)=e^{-\Delta_{\mathbb{W}}(W)}$, $\lambda_1(T_W)=e^{-\Delta_{\mathbb{W}}(W)}$) and from relation (4.5)

of Baraud and Birgé (2011), that there exist an at most countable collection $\mathbb{V}$ of models and a non-negative map Δ on $\mathbb{V}$ such that $\sum_{\mathbb{V}\in\mathbb{V}} e^{-\Delta(V)} \leq 1$ and

$$\begin{aligned} C \inf_{V\in\mathbb{V}} \left\{ d^2\left(\sqrt{s}|_A, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\} \\ \leq \inf_{W\in\mathbb{W}} \left\{ |\Phi|^2_{\infty,\sigma\wedge 1} d_2^{2(\sigma\wedge 1)}(u, T_W) + (\dim T_W + \Delta_{\mathbb{W}}(W))\tau_n \right\} \\ + \inf_{W\in\mathbb{W}} \left\{ d^2_\infty(\Phi, W) + (\dim W + \Delta_{\mathbb{W}}(W))\frac{\log n}{n} \right\} \end{aligned}$$

where $C > 0$ depends only on σ and where

$$\tau_n = \left(\log n \vee \log\left(|\Phi|_{\infty,\sigma\wedge 1}\right)\right)\frac{\log n}{n}.$$

Besides, for all linear space $V \in \mathbb{V}$, there exists a function $\psi \in \cup_{T\in\mathbb{T}_1} T$ and a linear space $W \in \mathbb{W}$ such that $V = \{f \circ \psi,\ f \in W\}$.

We apply Theorem 5 to $(\mathbb{V}, \Delta)$ to construct an estimator $\hat{s}$ of the form

$$\sqrt{\hat{s}(x,y)} = \hat{\phi}(y - \hat{g}(x))$$

such that

$$\begin{aligned} C' \inf_{V\in\mathbb{V}} \mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \inf_{W\in\mathbb{W}} \left\{ |\Phi|^2_{\infty,\sigma\wedge 1} d_2^{2(\sigma\wedge 1)}(u, T_W) + (\dim T_W + \Delta_{\mathbb{W}}(W))\tau_n \right\} \\ + \inf_{W\in\mathbb{W}} \left\{ d^2_\infty(\Phi, W) + (\dim W + \Delta_{\mathbb{W}}(W))\frac{\log n}{n} \right\} \end{aligned}$$

where $C' > 0$ depends only on σ, κ. We upper bound the two terms of the right-hand side of this inequality. We derive from

$$|\Phi|_{\infty,\sigma\wedge 1} \leq (1 + \|g\|_\infty)^{\sigma\wedge 1}|\phi|_{\infty,\sigma\wedge 1} \quad \text{and} \quad d_2^{2(\sigma\wedge 1)}(u, T_W) \leq \frac{d_2^{2(\sigma\wedge 1)}(g, W)}{(1 + \|g\|_\infty)^{2(\sigma\wedge 1)}}$$

that

$$\begin{aligned} \inf_{W\in\mathbb{W}} \left\{ |\Phi|^2_{\infty,\sigma\wedge 1} d_2^{2(\sigma\wedge 1)}(u, T_W) + (\dim T_W + \Delta_{\mathbb{W}}(W))\tau_n \right\} \\ \leq \inf_{W\in\mathbb{W}} \left\{ |\phi|^2_{\infty,\sigma\wedge 1} d_2^{2(\sigma\wedge 1)}(g, W) + (\dim W + \Delta_{\mathbb{W}}(W) + 1)\tau_n \right\}. \end{aligned}$$

By using the above lemma,

$$\begin{aligned} C'' \inf_{W\in\mathbb{W}} \left\{ |\phi|^2_{\infty,\sigma\wedge 1} d_2^{2(\sigma\wedge 1)}(g, W) + (\dim W + \Delta_{\mathbb{W}}(W) + 1)\tau_n \right\} \\ \leq \left(|\phi|_{\infty,\sigma\wedge 1}|g|^{\sigma\wedge 1}_{p,\beta}\right)^{\frac{2}{2(\sigma\wedge 1)\beta+1}} \tau_n^{\frac{2(\sigma\wedge 1)\beta}{2(\sigma\wedge 1)\beta+1}} + \tau_n \leq C''' \left(\frac{\log^2 n}{n}\right)^{\frac{2(\sigma\wedge 1)\beta}{2(\sigma\wedge 1)\beta+1}}. \end{aligned}$$

Similarly,

$$\begin{aligned} C'''' \inf_{W\in\mathbb{W}} \left\{ d^2_\infty(\Phi, W) + (\dim W + \Delta_{\mathbb{W}}(W))\frac{\log n}{n} \right\} \leq (|\Phi|_{\infty,\sigma})^{\frac{2}{2\sigma+1}} \left(\frac{\log n}{n}\right)^{\frac{2\sigma}{2\sigma+1}} + \frac{\log n}{n} \\ \leq C''''' \left(\frac{\log n}{n}\right)^{\frac{2\sigma}{2\sigma+1}}. \end{aligned}$$

□

6.9. Proof of Lemma 1. The first part of the lemma may be deduced from Proposition 4 of Baraud and Birgé (2011). For the second part, we shall build $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]} \notin \cup_{b>\sigma}\mathcal{H}^b([0,1])$ and $g' \in \mathcal{H}^\beta([0,1])$ such that $g'(0) = 0$ and

$$\phi' \circ g' \in \mathcal{H}^{\theta(\beta,\sigma)}([0,1]) \setminus \cup_{b>\theta(\beta,\sigma)}\mathcal{H}^b([0,1]).$$

By setting $\phi = \phi'$ and $g = -g'$, the function f defined by

$$f(x,y) = \phi'\left(y - (-g'(x))\right) \quad \text{for all } x, y \in [0,1],$$

is suitable since $f(x,0) = \phi' \circ g'(x)$ and $f(0,y) = \phi'(y)$.

If $\sigma, \beta \leq 1$, we can choose $\phi'(x) = x^\sigma$ on $[0,1]$ and $g'(x) = x^\beta$. If $\beta \geq \sigma \vee 1$, then choose $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]} \notin \cup_{b>\sigma}\mathcal{H}^b([0,1])$ and $g'(x) = x$. If now, $\sigma \geq \beta \vee 1$, we choose $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]} \notin \cup_{b>\sigma}\mathcal{H}^b([0,1])$ and such that $\phi'(x) = x$ for all $x \in [0,1/2]$. We then consider $\zeta \in \mathcal{H}^\beta([0,1]) \setminus \cup_{b>\beta}\mathcal{H}^b([0,1])$ and $g'(x) = (\zeta(x) - \zeta(0))/(2\sup_{y\in[0,1]}|\zeta(y) - \zeta(0)|)$. □

6.10. Proof of Corollary 5. We shall use the distances d_2 and d_∞ that have been defined at the beginning of the proof of Corollary 4.

Let us define

$$\forall x,y,z \in [0,1], \qquad u(x,y) = (u_1(x,y), u_2(x,y), u_3(x,y)) = \left(\frac{y - v_1(x)}{1 + \|v_1\|_\infty}, \frac{v_2(x)}{\|v_2\|_\infty}, \frac{v_3(x)}{\|v_3\|_\infty}\right)$$
$$\Phi(x,y,z) = \|v_3\|_\infty z\varphi\left((1 + \|v_1\|_\infty)\|v_2\|_\infty xy\right).$$

Let $\mathbb{W}$ be the family of linear spaces given by Lemma 8. Let for all $a \in \mathbb{R}$, $f \in \cup_{W\in\mathbb{W}}W$, $\psi_{a,f}$ be the function defined on $[0,1]^2$ by $\psi_{a,f}(x,y) = a(y - f(x))$ and g_f be the function defined on $[0,1]^3$ by $g_f(x,y,z) = zf(xy)$. For all $W \in \mathbb{W}$, we consider the linear spaces

$$T_W = \{\psi_{a,f},\, a \in \mathbb{R},\, f \in W\} \quad \text{and} \quad F_W = \{g_f,\, f \in W\}.$$

It ensues from the proof of Theorem 2 of Baraud and Birgé (2011) (where $l = 3$, $\mathbb{F} = \{F_W,\, W \in \mathbb{W}\}$, $\mathbb{T}_1 = \{T_W,\, W \in \mathbb{W}\}$, $\mathbb{T}_2 = \mathbb{T}_3 = \mathbb{W}$, $\gamma(F_W) = \lambda_1(T_W) = \lambda_2(W) = \lambda_3(W) = e^{-\Delta_{\mathbb{W}}(W)}$) that there exist an at most countable collection $\mathbb{V}$ of models and a non-negative map Δ on $\mathbb{V}$ such that $\sum_{\mathbb{V}\in\mathbb{V}} e^{-\Delta(V)} \leq 1$ and such that

$$\begin{aligned}
C \inf_{V\in\mathbb{V}} &\left\{ d^2\left(\sqrt{s}|_A, V\right) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\} \\
\leq\ & \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma\wedge 1)} \|v_2\|_\infty^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 d_2^{2(\sigma\wedge 1)}(u_1, T_W) + (\dim T_W + \Delta_{\mathbb{W}}(W))\tau_n^{(1)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 (1 + \|v_1\|_\infty)^{2(\sigma\wedge 1)} \|v_2\|_\infty^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 d_2^{2(\sigma\wedge 1)}(u_2, W) + (\dim W + \Delta_{\mathbb{W}}(W))\tau_n^{(1)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 \|\varphi\|_\infty^2 d_2^2(u_3, W) + (\dim W + \Delta_{\mathbb{W}}(W))\tau_n^{(2)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ d_\infty^2(\Phi, F_W) + (\dim F_W + \Delta_{\mathbb{W}}(W))\frac{\log n}{n} \right\}
\end{aligned}$$

where

$$\begin{aligned}\tau_n^{(1)} &= \left(\log n \vee \log\left(\|v_3\|_\infty^2(1+\|v_1\|_\infty)^{2(\sigma\wedge 1)}|\varphi|_{\infty,\sigma}^2\|v_2\|_\infty^{2(\sigma\wedge 1)}\right)\right)\frac{\log n}{n}\\ \tau_n^{(2)} &= \left(\log n \vee \log\left(\|v_3\|_\infty^2\|\varphi\|_\infty^2\right)\right)\frac{\log n}{n}.\end{aligned}$$

By applying Theorem 5 to $(\mathbb{V},\Delta)$, we build an estimator $\hat{s}$ such that

$$C'\mathbb{E}\left[H^2\left(s,\hat{s}\right)\right] \le \inf_{V\in\mathbb{V}}\left\{d^2\left(\sqrt{s}|_A, V\right) + \frac{\Delta(V)+\dim(V)\log n}{n}\right\}$$

and thus

$$\begin{aligned}C''\mathbb{E}\left[H^2\left(s,\hat{s}\right)\right] \le & \inf_{W\in\mathbb{W}}\left\{\|v_3\|_\infty^2|\varphi|_{\infty,\sigma}^2\|v_2\|_\infty^{2(\sigma\wedge 1)}d_2^{2(\sigma\wedge 1)}(v_1,W)+(\dim W+\Delta_{\mathbb{W}}(W))\tau_n^{(1)}\right\}\\ &+\inf_{W\in\mathbb{W}}\left\{\|v_3\|_\infty^2(1+\|v_1\|_\infty)^{2(\sigma\wedge 1)}|\varphi|_{\infty,\sigma}^2 d_2^{2(1\wedge\sigma)}(v_2,W)+(\dim W+\Delta_{\mathbb{W}}(W))\tau_n^{(1)}\right\}\\ &+\inf_{W\in\mathbb{W}}\left\{\|\varphi\|_\infty^2 d_2^{2(1\wedge\sigma)}(v_3,W)+(\dim W+\Delta_{\mathbb{W}}(W))\tau_n^{(2)}\right\}\\ &+\inf_{W\in\mathbb{W}}\left\{\|v_3\|_\infty^2 d_\infty^2(\varphi,W)+(\dim W+\Delta_{\mathbb{W}}(W))\frac{\log n}{n}\right\}.\end{aligned}$$

We conclude by applying Lemma 8 as in the end of the proof of Corollary 4. □

6.11. Proof of Lemma 2. The first part of the lemma can be deduced from Proposition 4 of Baraud and Birgé (2011). For the second part, remark that, as in the proof of Lemma 1 the problem amounts to finding $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$ with $\phi'|_{[0,1]} \notin \cup_{a>\sigma}\mathcal{H}^a(\mathbb{R})$, $v_i' \in \mathcal{H}^{\beta_i}([0,1])$ for $i\in\{1,2\}$, $v_1'(0)=0$, $v_2'(0)=1$ such that

$$\sqrt{v_2'}\,\phi'(v_1'v_2') \in \mathcal{H}^{\theta(\beta_1,\beta_2,\sigma)}([0,1]) \setminus \bigcup_{b>\theta(\beta_1,\beta_2,\sigma)}\mathcal{H}^b([0,1]).$$

If $\theta(\beta_1,\beta_2,\sigma)=2^{-1}(\beta_2\wedge 1)$, choose $v_2'(x)=(1-x)^{1\wedge\beta_2}$ and take ϕ' as being any function of $\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]}\notin\cup_{a>\sigma}\mathcal{H}^a(\mathbb{R})$ and such that $\phi'(0)=1$. If $\theta(\beta_1,\beta_2,\sigma)=\sigma$, choose $v_1'(x)=2(\sqrt{1+x}-1)$, $v_2'(x)=1/2(\sqrt{1+x}+1)$ and take ϕ' as being any function of $\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]}\notin\cup_{a>\sigma}\mathcal{H}^a(\mathbb{R})$. If $\theta(\beta_1,\beta_2,\sigma)=\sigma\beta_1$, we may assume that $\sigma\le 1$ and $\beta_1\le 1$. We can then choose $v_1'(x)=x^{\beta_1}$, $v_2'(x)=1$ and $\phi'(x)=x^\sigma$ for $x\in[0,1]$. If $\theta(\beta_1,\beta_2,\sigma)=\sigma\beta_2$, we may assume that $\sigma\le 1$ and $\beta_2\le 1$ and choose $v_1'(x)=1$ for $x\in[1/2,1]$, $v_2'(x)=1-(1-x)^{\beta_2}$ for $x\in[1/2,1]$ and $\phi'(x)=(1-x)^\sigma$ for $x\in[0,1]$. Finally, if $\theta(\beta_1,\beta_2,\sigma)=\beta_1$, we may assume that $\beta_1\le 1$. We can then choose $v_1'(x)=x^{\beta_1}$, $v_2'(x)=(1-x)^{1\wedge\beta_2}$ and ϕ' such that $\phi'(x)=x$ for $x\in[0,1/2]$. □

6.12. Proof of Proposition 6. We proceed in 3 steps.

Step 1. We associate to each cube $K\in\cup_{m\in\mathcal{M}_\ell}m$, a place in the computer's memory. Then, for each $i\in\{1,\dots,n\}$ we determine the sets $K\in\cup_{m\in\mathcal{M}_\ell}m$ such that $\mathbb{1}_K(X_i,X_{i+1})>0$.

There are at most ℓ such sets. This permits to store all the $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})$ in around $\mathcal{O}(n\ell d)$ operations. Let for all $K \in \cup_{m\in\mathcal{M}_\ell} m$, I_K and J_K be the subsets of $[0,1]^d$ such that $K = I_K \times J_K$. We can store all the $\mu(J_K)$ in $\mathcal{O}(4^{\ell d})$ operations and all the $\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)$ in $\mathcal{O}(n\ell d)$ operations. This permits us to store quickly

$$\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) \quad \text{and} \quad \sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x)\, \mathrm{d}\mu(x)$$

for all $K \in \cup_{m\in\mathcal{M}_\ell} m$. These values have to be calculated to know the $F_K(K')$ and thus to use the algorithm presented in Section 5.

Step 2. For each $K \in \cup_{m\in\mathcal{M}_\ell} m$, we use the algorithm of Section 5 to design m'_K. Let us denote by $j \in \{0, \dots, \ell\}$ the smallest integer such that $K \in \mathcal{K}_j$ where $\mathcal{K}_j$ is defined in Section 2.2.

- To find m'_K, we begin to compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \cup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_{\ell-1}} m$ such that $K'' \cap K \neq \emptyset$. The complexity of this is around the number of such sets, *i.e*, $4^{(\ell-j)d}$.
- Next, thanks to relation (11) we compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \cup_{m\in\mathcal{M}_{\ell-1}\setminus\mathcal{M}_{\ell-2}} m$ such that $K'' \cap K \neq \emptyset$. There are $4^{(\ell-j-1)d}$ such sets. The complexity of this operation is thus $4^d \times 4^{(\ell-j-1)d}$.
- By recurrence, we compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \cup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_j} m$ such that $K'' \cap K \neq \emptyset$ in at most
$$4^{(\ell-j)d} + 4^d \times \sum_{k=1}^{\ell-j-1} 4^{kd} \leq 3 \times 4^{(\ell-j)d}$$
operations.
- We get then $\mathcal{E}(T^\star([0,1]^d))$ in $4^d j$ additional operations.

We apply this algorithm for all $K \in \cup_{m\in\mathcal{M}_\ell} m$. When $K \in \mathcal{K}_j$, computing m'_K requires thus $\mathcal{O}\left(4^{(\ell-j)d} + 4^d j\right)$ operations. Since $|\mathcal{K}_j| = 4^{jd}$, computing all the m'_K requires finally

$$\sum_{j=0}^{\ell} 4^{jd} \left(4^{(\ell-j)d} + 4^d j\right) = \mathcal{O}\left(\ell 4^{(\ell+1)d}\right)$$

operations.

Step 3. Now, by slightly modifying the algorithm, we can compute (9) in $\mathcal{O}\left(4^{(\ell+1)d}\right)$ operations.

□

References

Akakpo, N. (2009). *Estimation adaptative par sélection de partitions en rectangles dyadiques.* PhD thesis, Université Paris Sud.

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21:1–28.

Akakpo, N. and Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653.

Athreya, K. B. and Atuncar, G. S. (1998). Kernel estimation for real-valued Markov chains. *Sankhyā. The Indian Journal of Statistics. Series A*, 60:1–17.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.

Baraud, Y. and Birgé, L. (2011). Estimating composite functions by model selection. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. To appear.

Basu, A. K. and Sahoo, D. K. (1998). On Berry-Esseen theorem for nonparametric density estimation in Markov sequences. *Bull. Inform. Cybernet.*, 30(1):25–39.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Probability Theory and Related Fields*, 65:181–237.

Birgé, L. (1984a). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, 20:201–223.

Birgé, L. (1984b). Sur un théorème de minimax et son application aux tests. *Probability and Mathematical Statistics*, 2:259–282.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2012). Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, volume 9, pages 47–64. IMS Collections.

Birgé, L. (2013). Model selection for density estimation with $\mathbb{L}_2$-loss. *Probability Theory and Related Fields*, pages 1–42.

Blanchard, G., Schäfer, C., and Rozenholc, Y. (2004). *Oracle bounds and exact algorithm for dyadic classification trees*, volume 3120 of *Lecture Notes in Comput. Sci.* Springer, Berlin.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144.

Clémencon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Mathematical Methods of Statistics*, 9:323–357.

Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1):111–145.

Dahmen, W., DeVore, R., and Scherer, K. (1980). Multi-dimensional spline approximation. *SIAM Journal on Numerical Analysis*, 17(3):pp. 380–402.

DeVore, R. and Yu, X. (1990). Degree of adaptive approximation. *Mathematics of Computation*, 55:625–635.

Dorea, C. C. Y. (2002). Strong consistency of kernel estimators for Markov transition densities. *Bull. Braz. Math. Soc. (N.S.)*, 33(3):409–418. Fifth Brazilian School in Probability (Ubatuba, 2001).

Doukhan, P. (1994). *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag. Properties and examples.

Doukhan, P. and Ghindès, M. (1983). Estimation de la transition de probabilité d'une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d'ordre 1. *Stochastic Process. Appl.*, 15(3):271–293.

Hochmuth, R. (2002). Wavelet characterizations for anisotropic besov spaces. *Applied and Computational Harmonic Analysis*, 12:179–208.

Juditsky, A., Lepski, O., and Tsybakov, A. (2009). Nonparametric estimation of composite functions. *The Annals of Statistics*, 37(3):1360–1404.

Lacour, C. (2007). Adaptive estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 43:571–597.

Lacour, C. (2008). Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications*, 118:232–260.

Lacour, C. (2012). Erratum to "Nonparametric estimation of the stationary density and the transition density of a Markov chain" [Stoch. Process. Appl. 118 (2008) 232–260] [mr2376901]. *Stochastic Process. Appl.*, 122(6):2480–2485.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54. Academic Press, New York.

Massart, P. (2003). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin/Heidelberg. École d'été de Probabilités de Saint-Flour.

Roussas, G. G. (1969). Nonparametric estimation in Markov processes. *Annals of the Institute of Statistical Mathematics*, 21:73–87.

Roussas, G. G. (1991). *Estimation of transition distribution function and its quantiles in Markov processes: strong consistency and asymptotic normality*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* Kluwer Acad. Publ., Dordrecht.

Viennet, G. (1997). Inequalities for absolutely regular sequences: application to density estimation. *Probability Theory and Related Fields*, 107:467–492.

CHAPITRE 4

Robust estimation on a parametric model with tests

Abstract

We are interested in the problem of robust parametric estimation of a density from i.i.d observations. By using a practice-oriented procedure based on robust tests, we build an estimator for which we establish non-asymptotic risk bounds with respect to the Hellinger distance under mild assumptions on the parametric model. We prove that the estimator is robust even for models for which the maximum likelihood method is bound to fail. We also evaluate the performance of the estimator by carrying out numerical simulations in which we observe that the estimator is very close to the maximum likelihood one when the model is regular enough and contains the true underlying density.

1. Introduction

Consider n independent and identically random variables $X_1, \ldots, X_n$ defined on an abstract probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with values in the measured space $(\mathbb{X}, \mathcal{F}, \mu)$. We suppose that the distribution of X_i admits a density s with respect to μ and aim at estimating s by using a parametric approach.

When the unknown density s is assumed to belong to a parametric model $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$ of densities, a traditional method to estimate $s = f_{\theta_0}$ is the maximum likelihood one. It is indeed well known that the maximum likelihood estimator (m.l.e for short) possesses nice statistical properties such as consistency and asymptotic efficiency when the model $\mathscr{F}$ is regular enough. However, it is also well known that this estimator breaks down for many models $\mathscr{F}$ of interest and counter examples may be found in Pitman (1979); Ferguson (1982); Le Cam (1990); Birgé (2006) among other references.

Another drawback of the m.l.e lies in the fact that it is not robust. This means that if s lies in a small neighbourhood of the model $\mathscr{F}$ but not in it, the m.l.e may perform poorly. Several kinds of robust estimators have been suggested in the literature to overcome this issue. We can cite the well known L and M estimators (which includes the class of minimum divergences estimators of Basu et al. (1998)) and the class of estimators built from a preliminary non-parametric estimator (such as the minimum Hellinger distance estimators introduced in Beran

(1977) and the related estimators of Lindsay (1994); Basu and Lindsay (1994)).

In this work, we focus on estimators built from robust tests. This approach, which begins in the 1970s with the works of Lucien Lecam and Lucien Birgé (Le Cam (1973, 1975); Birgé (1983, 1984a,b)), has the nice theoretical property to yield robust estimators under weak assumptions on the model $\mathscr{F}$. A key modern reference on this topic is Birgé (2006). The recent papers Birgé (2004, 2007, 2012, 2013); Baraud and Birgé (2009); Baraud (2011, 2013) show that increasing attention is being paid to this kind of estimator. Their main interest is to provide general theoretical results in various statistical settings (such as general model selection theorems) which are usually unattainable by the traditional procedures (such as those based on the minimization of a penalized contrast).

For our statistical issue, the procedures using tests are based on the pairwise comparison of the elements of a thin discretisation $\mathscr{F}_{\rm dis}$ of $\mathscr{F}$, that is, a finite or countable subset $\mathscr{F}_{\rm dis}$ of $\mathscr{F}$ such that for all function $f \in \mathscr{F}$, the distance between f and $\mathscr{F}_{\rm dis}$ is small (in a suitable sense). As a result, their complexities are of order the square of the cardinality of $\mathscr{F}_{\rm dis}$. Unfortunately, this cardinality is often very large, making the construction of the estimators difficult in practice. The aim of this chapter is to develop a faster way of using tests to build an estimator when the cardinality of $\mathscr{F}_{\rm dis}$ is large.

From a theoretical point of view, the estimator we propose possesses similar statistical properties than those proved in Birgé (2006); Baraud (2011). Under mild assumptions on $\mathscr{F}$, we build an estimator $\hat{s} = f_{\hat{\theta}}$ of s such that

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \geq \inf_{\theta\in\Theta} h^2(s, f_\theta) + \frac{d}{n} + \xi\right] \leq e^{-n\xi} \quad \text{for all } \xi > 0, \tag{1}$$

where C is a positive number depending on $\mathscr{F}$, h the Hellinger distance and d such that $\Theta \subset \mathbb{R}^d$. We recall that the Hellinger distance is defined on the cone $\mathbb{L}_+^1(\mathbb{X}, \mu)$ of non-negative integrable functions on $\mathbb{X}$ with respect to μ by

$$h^2(f, g) = \frac{1}{2}\int_{\mathbb{X}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}\mu(x) \quad \text{for all } f, g \in \mathbb{L}_+^1(\mathbb{X}, \mu).$$

Let us make some comments on (1). When s does belong to the model $\mathscr{F}$, the estimator achieves a quadratic risk of order n^{-1} with respect to the Hellinger distance. Besides, there exists $\theta_0 \in \Theta$ such that $s = f_{\theta_0}$ and we may then derive from (1) the rate of convergence of $\hat{\theta}$ to θ_0. In general, we do not suppose that the unknown density belongs to the model but rather use $\mathscr{F}$ as an approximate class (sieve) for s. Inequality (1) shows then that the estimator $\hat{s} = f_{\hat{\theta}}$ cannot be strongly influenced by small departures from the model. As a matter of fact, if $\inf_{\theta\in\Theta} h^2(s, f_\theta) \leq n^{-1}$, which means that the model is slightly misspecified, the quadratic risk of the estimator $\hat{s} = f_{\hat{\theta}}$ remains of order n^{-1}. This can be interpreted as a robustness property.

The preceding inequality (1) is interesting because it proves that our estimator is robust and converges at the right rate of convergence when the model is correct. However, the constant C depends on several parameters on the model such as the size of Θ. It is thus far from obvious that such an estimator can be competitive against more traditional estimators (such as the m.l.e).

In this work, we try to give a partial answer for our estimator by carrying out numerical simulations. When a very thin discretisation $\mathscr{F}_{\rm dis}$ is used, the simulations show that our estimator

is very close to the m.l.e when the model is regular enough and contains s. More precisely, the larger is the number of observations n, the closer they are, suggesting that our estimator inherits the efficiency of the m.l.e. Of course, this does not in itself constitute a proof but this allows to indicate what kind of results can be expected. A theoretical connection between estimators built from tests (with the procedure described in Baraud (2011)) and the m.l.e will be found in a future paper of Yannick Baraud and Lucien Birgé.

In this chapter, we consider the problem of estimation on a single model. Nevertheless, when the statistician has at disposal several candidate models for s, a natural issue is model selection. In order to address it, one may associate to each of these models the estimator resulting from our procedure and then select among those estimators by means of the procedure of Baraud (2011). By combining his Theorem 2 with our risk bounds on each individual estimator, we obtain that the selected estimator satisfies an oracle-type inequality.

We organize this chapter as follows. We begin with a glimpse of the results in Section 2. We then present a procedure and its associated theoretical results to deal with models parametrized by an unidimensional parameter in Section 3. We evaluate its performance in practice by carrying out numerical simulations in Section 4. We work with models parametrized by a multidimensional parameter in Sections 5 and 6. The proofs are postponed to Section 6. Some technical results about the practical implementation of our procedure devoted to the multidimensional models are delayed to Section 7.

Let us introduce some notations that will be used all along the chapter. The number $x \vee y$ (respectively $x \wedge y$) stands for $\max(x, y)$ (respectively $\min(x, y)$) and x_+ stands for $x \vee 0$. We set $\mathbb{N}^\star = \mathbb{N} \setminus \{0\}$. The vector $(\theta_1, \ldots, \theta_d)$ of $\mathbb{R}^d$ is denoted by the bold letter $\boldsymbol{\theta}$. Given a set of densities $\mathscr{F} = \{f_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$, for all $A \subset \Theta$, the notation diam$A$ stands for $\sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in A} h^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})$. The cardinality of a finite set A is denoted by $|A|$. For (E, d) a metric space, $x \in E$ and $A \subset E$, the distance between x and A is denoted by $d(x, A) = \inf_{a \in A} d(x, a)$. The indicator function of a subset A is denoted by $\mathbb{1}_A$. The notations C,C',C''... are for the constants. The constants C,C',C''... may change from line to line.

2. An overview of the chapter

2.1. Assumption. In this chapter, we shall deal with sets of densities $\mathscr{F} = \{f_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$ indexed by a rectangle

$$\Theta = \prod_{j=1}^{d} [m_j, M_j]$$

of $\mathbb{R}^d$. A such set will be called model. From now on, we consider models satisfying the following assumption.

Assumption 1. *There exist positive numbers $\alpha_1, \ldots, \alpha_d$, $\underline{R}_1, \ldots, \underline{R}_d$, $\overline{R}_1, \ldots, \overline{R}_d$ such that for all $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $\boldsymbol{\theta}' = (\theta'_1, \ldots, \theta'_d) \in \Theta = \prod_{j=1}^d [m_j, M_j]$*

$$\sup_{j \in \{1,\ldots,d\}} \underline{R}_j |\theta_j - \theta'_j|^{\alpha_j} \leq h^2\left(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}\right) \leq \sup_{j \in \{1,\ldots,d\}} \overline{R}_j |\theta_j - \theta'_j|^{\alpha_j}.$$

This assumption allows to connect a (quasi) distance between the parameters to the Hellinger one between the corresponding densities. A similar assumption may be found in Theorem 5.8 of Chapter 1 of Ibragimov and Has'minskii (1981) to prove results on the maximum likelihood estimator. They require however that the application $\theta \mapsto f_\theta(x)$ is continuous for μ almost all x to ensure the existence and the consistency of the m.l.e. Without this additional assumption, the m.l.e may not exist as shown by the translation model

$$\mathscr{F} = \{f_\theta,\, \theta \in [-1,1]\} \quad \text{where} \quad f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}}\mathbb{1}_{[-1,1]}(x-\theta) & \text{for all } x \in \mathbb{R}\setminus\{\theta\} \\ 0 & \text{for } x = \theta \end{cases}$$

for which Assumption 1 holds with $\alpha_1 = 1/2$.

Under suitable regularity conditions on the model, Theorem 7.6 of Chapter 1 of Ibragimov and Has'minskii (1981) shows that this assumption is fulfilled with $\alpha_1 = \cdots = \alpha_d = 2$. Other kinds of sufficient conditions implying Assumption 1 may be found in this book (see the beginning of Chapter 5 and Theorem 1.1 of Chapter 6). Other examples and counter-examples are given in Chapter 7 of Dacunha-Castelle (1978). Several models of interest satisfying this assumption will appear later in the chapter.

2.2. Risk bound. In this chapter, the risk bound we get for our estimator $\hat{s}$ is similar to the one we would get by the procedures of Birgé (2006); Baraud (2011). More precisely:

Theorem 1. *Suppose that Assumption 1 holds. We can build an estimator $\hat{s}$ of the form $\hat{s} = f_{\hat{\boldsymbol{\theta}}}$ such that for all $\xi > 0$.*

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\boldsymbol{\theta}}}) \geq h^2(s,\mathscr{F}) + \frac{d}{n} + \xi\right] \leq e^{-n\xi} \tag{2}$$

where $C > 0$ depends on $\sup_{1\leq j\leq d} \overline{R}_j/\underline{R}_j$ and $\min_{1\leq j\leq d}\alpha_j$.

We deduce from this risk bound that if $s = f_{\boldsymbol{\theta}_0}$ belongs to the model $\mathscr{F}$, the estimator $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and the random variable $h^2(s, f_{\hat{\boldsymbol{\theta}}})$ is of order n^{-1}. Besides, we may then derive from Assumption 1 that there exist positive numbers a, b_j such that

$$\mathbb{P}\left[n^{1/\alpha_j}\left|\hat{\theta}_j - \theta_{0,j}\right| \geq \xi\right] \leq ae^{-b_j\xi^{\alpha_j}} \quad \text{for all } j \in \{1,\ldots,d\} \text{ and } \xi > 0.$$

Precisely, $a = e^d$ and $b_j = C\underline{R}_j$. We emphasize here that this exponential inequality on $\hat{\theta}_j$ is non-asymptotic but that the constants a, b_j are unfortunately far from optimal.

As explained in the introduction, there is no assumption on the true underlying density s, which means that the model $\mathscr{F}$ may be misspecified. In particular, when the squared Hellinger distance between the unknown density and the model $\mathscr{F}$ is of order n^{-1}, the random variable $h^2(s, f_{\hat{\boldsymbol{\theta}}})$ remains of order n^{-1}. This shows that the estimator $\hat{s}$ possesses robustness properties.

2.3. Numerical complexity. The main interest of our procedures with respect to those of Birgé (2006); Baraud (2011) lies in their numerical complexity. More precisely, we shall prove the proposition below.

Proposition 2. *Under Assumption 1, we can build an estimator $\hat{s}$ satisfying (2) by computing less than*

$$4C^{d/\bar{\boldsymbol{\alpha}}}\left[\prod_{j=1}^{d}\left(1+\left(\overline{R}_j/\underline{R}_j\right)^{1/\alpha_j}\right)\right]\left[\sum_{j=1}^{d}\max\left\{1,\log\left((n\overline{R}_j/d)^{1/\alpha_j}(M_j-m_j)\right)\right\}\right]$$

tests. In the above inequality, C is a constant larger than 1 (independent of n and the model $\mathscr{F}$) and $\bar{\boldsymbol{\alpha}}$ stands for the harmonic mean of $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_d)$, that is

$$\frac{1}{\bar{\boldsymbol{\alpha}}}=\frac{1}{d}\sum_{j=1}^{d}\frac{1}{\alpha_j}.$$

If we are interested in the complexity when n is large, the number of tests computed is asymptotically equivalent to $C'\log n$ where

$$C'=4(d/\bar{\boldsymbol{\alpha}})C^{d/\bar{\boldsymbol{\alpha}}}\prod_{j=1}^{d}\left(1+\left(\overline{R}_j/\underline{R}_j\right)^{1/\alpha_j}\right).$$

It is worthwhile to notice that the number of tests computed grows slowly with n. The constant C' is not too large if d, $1/\bar{\boldsymbol{\alpha}}$, $(\overline{R}_j/\underline{R}_j)^{1/\alpha_j}$ are small enough.

Remark. The constant C' does not depend only on the model but also on its parametrisation. As a matter of fact, in the uniform model

$$\mathscr{F}=\{f_\theta,\,\theta\in[m_1,M_1]\}\quad\text{where}\quad f_\theta=\theta^{-1}\mathbb{1}_{[0,\theta]}$$

we can compute explicitly the Hellinger distance

$$h^2(f_\theta,f_{\theta'})=\frac{|\theta'-\theta|}{(\sqrt{\theta}+\sqrt{\theta'})\sqrt{\max(\theta,\theta')}}$$

and bound it from above and from below by

$$\frac{1}{2M_1}|\theta'-\theta|\le h^2(f_\theta,f_{\theta'})\le\frac{1}{2m_1}|\theta'-\theta|.$$

Now, if we parametrise $\mathscr{F}$ as

$$\mathscr{F}=\{f_{e^t},\,t\in[\log m_1,\log M_1]\},$$

then the Hellinger becomes $h^2(f_{e^t},f_{e^{t'}})=1-e^{-|t'-t|/2}$, and we can bound it from above and from below by

$$\frac{1-\sqrt{m_1/M_1}}{\log(M_1/m_1)}|t'-t|\le h^2(f_{e^t},f_{e^{t'}})\le\frac{1}{2}|t'-t|.$$

Assumption 1 is satisfied in both case but with different values of $\underline{R}_1$ and $\overline{R}_1$. When M_1/m_1 is large, the second parametrisation is much more interesting since it leads to a smaller constant C'.

3. Models parametrized by an unidimensional parameter

We now describe our procedure when the parametric model $\mathscr{F}$ is indexed by an interval $\Theta = [m_1, M_1]$ of $\mathbb{R}$. Throughout this section, Assumption 1 is supposed to be fulfilled. For the sake of simplicity, the subscripts of m_1, M_1 and α_1 are omitted.

3.1. Basic ideas. We begin to detail the heuristics on which is based our procedure. We assume in this section that s belongs to the model $\mathscr{F}$, that is, there exists $\theta_0 \in \Theta = [m, M]$ such that $s = f_{\theta_0}$. The starting point is the existence for all $\theta, \theta' \in \Theta$ of a measurable function $T(\theta, \theta')$ of the observations $X_1, \dots, X_n$ such that

1. For all $\theta, \theta' \in \Theta$, $T(\theta, \theta') = -T(\theta', \theta)$.

2. There exists $\kappa > 0$ such that if $\mathbb{E}[T(\theta, \theta')]$ is non-negative, then
$$h^2(s, f_\theta) > \kappa h^2(f_\theta, f_{\theta'}).$$

3. For all $\theta, \theta' \in \Theta$, $T(\theta, \theta')$ and $\mathbb{E}[T(\theta, \theta')]$ are close (in a suitable sense).

For all $\theta \in \Theta$, $r > 0$, let $\mathcal{B}(\theta, r)$ be the Hellinger ball centered at θ with radius r, that is
$$\mathcal{B}(\theta, r) = \left\{\theta' \in \Theta,\, h(f_\theta, f_{\theta'}) \leq r\right\}. \tag{3}$$
For all $\theta, \theta' \in \Theta$, we deduce from the first point that either $T(\theta, \theta')$ is non-negative, or $T(\theta', \theta)$ is non-negative. It is likely that it follows from 2. and 3. that in the first case
$$\theta_0 \in \Theta \setminus \mathcal{B}\big(\theta, \kappa^{1/2} h(f_\theta, f_{\theta'})\big)$$
while in the second case
$$\theta_0 \in \Theta \setminus \mathcal{B}\big(\theta', \kappa^{1/2} h(f_\theta, f_{\theta'})\big).$$
These sets may be interpreted as confidence sets for θ_0.

The main idea is to build a decreasing sequence (in the sense of inclusion) of intervals $(\Theta_i)_i$. Set $\theta^{(1)} = m$, $\theta'^{(1)} = M$, and $\Theta_1 = [\theta^{(1)}, \theta'^{(1)}]$ (which is merely Θ). If $T(\theta^{(1)}, \theta'^{(1)})$ is non-negative, we consider a set Θ_2 such that
$$\Theta_1 \setminus \mathcal{B}\left(\theta^{(1)}, \kappa^{1/2} h(f_{\theta^{(1)}}, f_{\theta'^{(1)}})\right) \subset \Theta_2 \subset \Theta_1$$
while if $T(\theta^{(1)}, \theta'^{(1)})$ is non-positive, we consider a set Θ_2 such that
$$\Theta_1 \setminus \mathcal{B}\left(\theta'^{(1)}, \kappa^{1/2} h(f_{\theta^{(1)}}, f_{\theta'^{(1)}})\right) \subset \Theta_2 \subset \Theta_1.$$
The set Θ_2 may thus also be interpreted as a confidence set for θ_0. Thanks to Assumption 1, we can define Θ_2 as an interval $\Theta_2 = [\theta^{(2)}, \theta'^{(2)}]$.

We then repeat the idea to build an interval $\Theta_3 = [\theta^{(3)}, \theta'^{(3)}]$ included in Θ_2 and containing either
$$\Theta_3 \supset \Theta_2 \setminus \mathcal{B}\left(\theta^{(2)}, \kappa^{1/2} h(f_{\theta^{(2)}}, f_{\theta'^{(2)}})\right) \quad \text{or} \quad \Theta_3 \supset \Theta_2 \setminus \mathcal{B}\left(\theta'^{(2)}, \kappa^{1/2} h(f_{\theta^{(2)}}, f_{\theta'^{(2)}})\right)$$

according to the sign of $T(\theta^{(2)}, \theta'^{(2)})$.

By induction, we build a decreasing sequence of such intervals $(\Theta_i)_i$. We now consider an integer N large enough so that the length of Θ_N is small enough. We then define the estimator $\hat{\theta}$ as the center of the set Θ_N and estimate s by $f_{\hat{\theta}}$.

3.2. Definition of the test. The test $T(\theta, \theta')$ we use in our estimation strategy is the one of Baraud (2011) applied to two suitable densities of the model. More precisely, let $\overline{T}$ be the functional defined for all $g, g' \in \mathbb{L}_+^1(\mathbb{X}, \mu)$ by

$$\overline{T}(g, g') = \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{g'(X_i)} - \sqrt{g(X_i)}}{\sqrt{g(X_i) + g'(X_i)}} + \frac{1}{2}\int_{\mathbb{X}} \sqrt{g(x) + g'(x)}\left(\sqrt{g'(x)} - \sqrt{g(x)}\right) \mathrm{d}\mu(x) \qquad (4)$$

where the convention $0/0 = 0$ is in use.

We consider $t \in (0, 1]$ and $\epsilon = t(\overline{R}n)^{-1/\alpha}$. We then define the finite sets

$$\Theta_{\mathrm{dis}} = \left\{m + k\epsilon,\ k \in \mathbb{N},\ k \leq (M-m)\epsilon^{-1}\right\}, \quad \mathscr{F}_{\mathrm{dis}} = \{f_\theta,\ \theta \in \Theta_{\mathrm{dis}}\}$$

and the map π on $[m, M]$ by

$$\pi(x) = m + \lfloor (x-m)/\epsilon \rfloor \epsilon \quad \text{for all } x \in [m, M]$$

where $\lfloor \cdot \rfloor$ denotes the integer part. We then define $T(\theta, \theta')$ by

$$T(\theta, \theta') = \overline{T}(f_{\pi(\theta)}, f_{\pi(\theta')}) \quad \text{for all } \theta, \theta' \in [m, M].$$

The aim of the parameter t is to tune the thinness of the net $\mathscr{F}_{\mathrm{dis}}$.

3.3. Procedure. We shall build a decreasing sequence $(\Theta_i)_{i\geq 1}$ of intervals of $\Theta = [m, M]$ as explained in Section 3.1. Let $\kappa > 0$, and for all $\theta, \theta' \in [m, M]$ such that $\theta' < \theta$, let $\bar{r}(\theta, \theta')$, $\underline{r}(\theta, \theta')$ be two positive numbers such that

$$[m, M] \bigcap \left[\theta, \theta + \bar{r}(\theta, \theta')\right] \quad \subset \quad \mathcal{B}\big(\theta, \kappa^{1/2} h(f_\theta, f_{\theta'})\big) \qquad (5)$$

$$[m, M] \bigcap \left[\theta' - \underline{r}(\theta, \theta'), \theta'\right] \quad \subset \quad \mathcal{B}\big(\theta', \kappa^{1/2} h(f_\theta, f_{\theta'})\big) \qquad (6)$$

where we recall that $\mathcal{B}(\theta, \kappa^{1/2}h(f_\theta, f_{\theta'}))$ and $\mathcal{B}(\theta', \kappa^{1/2}h(f_\theta, f_{\theta'}))$ are the Hellinger balls defined by (3).

We set $\theta^{(1)} = m$, $\theta'^{(1)} = M$ and $\Theta_1 = [\theta^{(1)}, \theta'^{(1)}]$. We define the sequence $(\Theta_i)_{i\geq 1}$ by induction. When $\Theta_i = [\theta^{(i)}, \theta'^{(i)}]$, we set

$$\theta^{(i+1)} = \begin{cases} \theta^{(i)} + \min\left\{\bar{r}(\theta^{(i)}, \theta'^{(i)}), \frac{\theta'^{(i)} - \theta^{(i)}}{2}\right\} & \text{if } T(\theta^{(i)}, \theta'^{(i)}) \geq 0 \\ \theta^{(i)} & \text{otherwise} \end{cases}$$

$$\theta'^{(i+1)} = \begin{cases} \theta'^{(i)} - \min\left\{\underline{r}(\theta^{(i)}, \theta'^{(i)}), \frac{\theta'^{(i)} - \theta^{(i)}}{2}\right\} & \text{if } T(\theta^{(i)}, \theta'^{(i)}) \leq 0 \\ \theta'^{(i)} & \text{otherwise.} \end{cases}$$

We then define $\Theta_{i+1} = [\theta^{(i+1)}, \theta'^{(i+1)}]$.

The role of conditions (5) and (6) is to ensure that Θ_{i+1} is big enough to contain one of the two confidence sets

$$\Theta_i \setminus \mathcal{B}\left(\theta^{(i)}, \kappa^{1/2} h(f_{\theta^{(i)}}, f_{\theta'^{(i)}})\right) \quad \text{and} \quad \Theta_i \setminus \mathcal{B}\left(\theta'^{(i)}, \kappa^{1/2} h(f_{\theta^{(i)}}, f_{\theta'^{(i)}})\right).$$

The parameter κ allows to tune the level of these confidence sets. There is a minimum in the definitions of $\theta^{(i+1)}$ and $\theta'^{(i+1)}$ in order to guarantee the inclusion of Θ_{i+1} in Θ_i.

We now consider a positive number η and build these intervals until their lengths become smaller than η. The estimator we consider is then the center of the last interval built. This parameter η stands for a measure of the accuracy of the estimation and must be small enough to get a suitable risk bound for our estimator.

The algorithm is the following.

Algorithm 1

$\theta \leftarrow m$, $\theta' \leftarrow M$
while $\theta' - \theta > \eta$ **do**
  Compute $r = \min\{\bar{r}(\theta, \theta'), (\theta' - \theta)/2\}$
  Compute $r' = \min\{\underline{r}(\theta, \theta'), (\theta' - \theta)/2\}$
  Compute Test $= T(\theta, \theta')$
  if Test ≥ 0 **then**
    $\theta \leftarrow \theta + r$
  end if
  if Test ≤ 0 **then**
    $\theta' \leftarrow \theta' - r'$
  end if
end while
Return: $\hat{\theta} = (\theta + \theta')/2$

3.4. Risk bound. The following theorem specify the values of the parameters t, κ, η that allow to control the risk of the estimator $\hat{s} = f_{\hat{\theta}}$.

Theorem 3. *Suppose that Assumption 1 holds. Set*

$$\bar{\kappa} = \left(1 + \sqrt{\frac{2+\sqrt{2}}{2-\sqrt{2}}}\right)^{-2}. \tag{7}$$

Assume that $t \in (0, 1]$, $\kappa \in (0, \bar{\kappa})$, $\eta \in [\epsilon, (\overline{R}n)^{-1/\alpha}]$ and that $\bar{r}(\theta, \theta')$, $\underline{r}(\theta, \theta')$ are such that (5) and (6) hold.

Then, for all $\xi > 0$, the estimator $\hat{\theta}$ built in Algorithm 1 satisfies

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \geq h^2(s, \mathscr{F}) + \frac{1}{n} + \xi\right] \leq e^{-n\xi}$$

where $C > 0$ depends only on $\kappa, t, \alpha, \overline{R}/\underline{R}$.

A slightly sharper risk bound may be found in the proof of this theorem.

3.5. Choice of $\bar{r}(\theta,\theta')$ and $\underline{r}(\theta,\theta')$. These parameters are chosen by the statistician. They do not change the risk bound given by Theorem 3 (provided that (5) and (6) hold) but affect the speed of the procedure. The larger they are, the faster the procedure is. There are three different situations.

First case: the Hellinger distance $h(f_\theta, f_{\theta'})$ can be made explicit. We have thus an interest in defining them as the largest numbers for which (5) and (6) hold, that is

$$\bar{r}(\theta,\theta') = \sup\left\{r>0,\ [m,M]\cap[\theta,\theta+r]\subset\mathcal{B}\big(\theta,\kappa^{1/2}h(f_\theta,f_{\theta'})\big)\right\} \tag{8}$$

$$\underline{r}(\theta,\theta') = \sup\left\{r>0,\ [m,M]\cap\big[\theta'-r,\theta'\big]\subset\mathcal{B}\big(\theta',\kappa^{1/2}h(f_\theta,f_{\theta'})\big)\right\}. \tag{9}$$

Second case: the Hellinger distance $h(f_\theta, f_{\theta'})$ can be quickly evaluated numerically but the computation of (8) and (9) is difficult. We may then define them by

$$\underline{r}(\theta,\theta') = \bar{r}(\theta,\theta') = \big((\kappa/\bar{R})h^2(f_\theta,f_{\theta'})\big)^{1/\alpha}. \tag{10}$$

One can verify that (5) and (6) hold. When the model is regular enough and $\alpha=2$, the value of $\overline{R}$ can be calculated by using Fisher information (see for instance Theorem 7.6 of Chapter 1 of Ibragimov and Has'minskii (1981)).

Third case: the computation of the Hellinger distance $h(f_\theta, f_{\theta'})$ involves the numerical computation of an integral and this computation is slow. An alternative definition is then

$$\underline{r}(\theta,\theta') = \bar{r}(\theta,\theta') = (\kappa\underline{R}/\overline{R})^{1/\alpha}\left(\theta'-\theta\right). \tag{11}$$

As in the second point, one can check that (5) and (6) hold. Note however that the computation of the test also involves in most cases the numerical computation of an integral (see (4)). This third case is thus mainly devoted to models for which this numerical integration can be avoided, as for the translation models $\mathscr{F}=\{f(\cdot-\theta),\ \theta\in[m,M]\}$ with f even, $\mathbb{X}=\mathbb{R}$ and μ the Lebesgue measure (the second term of (4) is 0 for these models).

We can upper-bound the numerical complexity of the algorithm when $\bar{r}(\theta,\theta')$ and $\underline{r}(\theta,\theta')$ are large enough. Precisely, we prove the proposition below.

Proposition 4. *Suppose that the assumptions of Theorem 3 hold and that $\underline{r}(\theta,\theta')$, $\bar{r}(\theta,\theta')$ are larger than*

$$(\kappa\underline{R}/\overline{R})^{1/\alpha}\left(\theta'-\theta\right). \tag{12}$$

Then, the number of tests computed to build the estimator $\hat{\theta}$ is smaller than

$$1+\max\left\{\left(\overline{R}/(\kappa\underline{R})\right)^{1/\alpha},1/\log 2\right\}\log\left(\frac{M-m}{\eta}\right).$$

It is worthwhile to notice that this upper-bound does not depend on t, that is the size of the net $\mathscr{F}_{\text{dis}}$ contrary to the preceding procedures based on tests. Obviously, the parameter η is involved in this upper-bound, but the whole point is that it grows slowly with $1/\eta$, which allows to use the procedure with η very small.

4. Simulations for unidimensional models

In what follows, we carry out a simulation study in order to evaluate more precisely the performance of our estimator. We simulate samples $(X_1, \dots, X_n)$ with density s and use our procedure to estimate s.

4.1. Models. Our simulation study is based on the following models.

Example 1. $\mathscr{F} = \{f_\theta,\ \theta \in [0.01, 100]\}$ *where*

$$f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{[0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

Example 2. $\mathscr{F} = \{f_\theta,\ \theta \in [-100, 100]\}$ *where*

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \quad \textit{for all } x \in \mathbb{R}.$$

Example 3. $\mathscr{F} = \{f_\theta,\ \theta \in [0.01, 100]\}$ *where*

$$f_\theta(x) = \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right) \mathbb{1}_{[0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

Example 4. $\mathscr{F} = \{f_\theta,\ \theta \in [-10, 10]\}$ *where*

$$f_\theta(x) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} \quad \textit{for all } x \in \mathbb{R}.$$

Example 5. $\mathscr{F} = \{f_\theta,\ \theta \in [0.01, 10]\}$ *where* $f_\theta = \theta^{-1} \mathbb{1}_{[0,\theta]}$.

Example 6. $\mathscr{F} = \{f_\theta,\ \theta \in [-10, 10]\}$ *where*

$$f_\theta(x) = \frac{1}{(x-\theta+1)^2} \mathbb{1}_{[\theta,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

Example 7. $\mathscr{F} = \{f_\theta,\ \theta \in [-10, 10]\}$ *where* $f_\theta = \mathbb{1}_{[\theta-1/2,\theta+1/2]}$.

Example 8. $\mathscr{F} = \{f_\theta,\ \theta \in [-1, 1]\}$ *where*

$$f_\theta(x) = \frac{1}{4\sqrt{|x-\theta|}} \mathbb{1}_{[-1,1]}(x-\theta) \quad \textit{for all } x \in \mathbb{R} \setminus \{\theta\}$$

and $f_\theta(\theta) = 0$.

In these examples, we shall mainly compare our estimator to the maximum likelihood one. In examples 1,2,3,5 and 6, the m.l.e $\tilde{\theta}_{\text{mle}}$ can be made explicit and is thus easy to compute. Finding the m.l.e is more delicate for the problem of estimating the location parameter of a Cauchy distribution, since the likelihood function may be multimodal. We refer to Barnett (1966) for a discussion of numerical methods devoted to the maximization of the likelihood. In our simulation study, we avoid the issues of the numerical algorithms by computing the likelihood at 10^6 equally spaced points between $\max(-10, \hat{\theta}-1)$ and $\min(10, \hat{\theta}+1)$ (where $\hat{\theta}$ is our estimator)

and at 10^6 equally spaced points between $\max(-10, \tilde{\theta}_{\text{median}} - 1)$ and $\min(10, \tilde{\theta}_{\text{median}} + 1)$ where $\tilde{\theta}_{\text{median}}$ is the median. We then select among these points the one for which the likelihood is maximal. In Example 5, we shall also compare our estimator to the minimum variance unbiased estimator defined by

$$\tilde{\theta}_{\text{mvub}} = \frac{n+1}{n} \max_{1 \le i \le n} X_i.$$

In Example 7, we shall compare our estimator to

$$\tilde{\theta}' = \frac{1}{2} \left(\max_{1 \le i \le n} X_i + \min_{1 \le i \le n} X_i \right).$$

In the case of Example 8, the likelihood is infinite at each observation and the maximum likelihood method fails. We shall then compare our estimator to the median and the empirical mean but also to the maximum spacing product estimator $\tilde{\theta}_{\text{mspe}}$ (m.s.p.e for short). This estimator was introduced by Cheng and Amin (1983); Ranneby (1984) to deal with statistical models for which the likelihood is unbounded. The m.s.p.e is known to possess nice theoretical properties such as consistency and asymptotic efficiency and precise results on the performance of this estimator may be found in Cheng and Amin (1983); Ranneby (1984); Ekström (1998); Shao and Hahn (1999); Ghost and Jammalamadaka (2001); Anatolyev and Kosenok (2005) among other references. This last method involves the problem of finding a global maximum of the maximum product function on $\Theta = [-1, 1]$. We compute it by considering 2×10^5 equally spaced points between -1 and 1 and by calculating for each of these points the function to maximize. We then select the point for which the function is maximal. Using more points to compute the m.s.p.e would give more accurate results, especially when n is large, but we are limited by the computer.

4.2. Implementation of the procedure. Our procedure involves several parameters that must be chosen by the statistician.

Choice of t. This parameter tunes the thinness of the net $\mathscr{F}_{\text{dis}}$. When the model is regular enough and contains s, a good choice of t seems to be $t = 0$ (that is $\Theta_{\text{dis}} = \Theta$, $\mathscr{F}_{\text{dis}} = \mathscr{F}$ and $T(\theta, \theta') = \overline{T}(f_\theta, f_{\theta'})$), since then the simulations suggest that our estimator is very close to the m.l.e when the model is true (with large probability). In the simulations, we take $t = 0$.

Choice of η. We take η small: $\eta = (M - m)/10^8$.

Choice of κ. This constant influences the level of the confidence sets and thus the time of construction of the estimator: the larger is κ, the faster is the procedure. We take arbitrary $\kappa = \bar{\kappa}/2$.

Choice of $\underline{r}(\theta, \theta')$ **and** $\bar{r}(\theta, \theta')$. In examples 1,2,3,5, and 7, we define them by (8) and (9). In examples 4 and 6, we define them by (10). In the first case, $\alpha = 2$ and $\overline{R} = 1/16$, while in the second case, $\alpha = 1$ and $\overline{R} = 1/2$. In the case of Example 8, we use (11) with $\alpha = 1/2$, $\underline{R} = 0.17$ and $\overline{R} = 1/\sqrt{2}$.

4.3. Simulations when $s \in \mathscr{F}$**.** We begin to simulate N samples $(X_1, \ldots, X_n)$ when the true density s belongs to the model $\mathscr{F}$. They are generated according to the density $s = f_1$ in examples $1, 3, 5$ and according to $s = f_0$ in examples $2, 4, 6, 7, 8$.

We evaluate the performance of an estimator $\tilde{\theta}$ by computing it on each of the N samples. Let $\tilde{\theta}^{(i)}$ be the value of this estimator corresponding to the i^{th} sample and let

$$\widehat{R}_N(\tilde{\theta}) = \frac{1}{N}\sum_{i=1}^{N} h^2(s, f_{\tilde{\theta}^{(i)}}) \quad \text{and} \quad \widehat{\text{std}}_N(\tilde{\theta}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(h^2(s, f_{\tilde{\theta}^{(i)}}) - \widehat{R}_N(\tilde{\theta})\right)^2}.$$

The risk $\mathbb{E}\left[h^2(s, f_{\tilde{\theta}})\right]$ of the estimator $\tilde{\theta}$ is thus estimated by $\widehat{R}_N(\tilde{\theta})$. More precisely, if Q_c denotes the $c/2$ quantile of a standard Gaussian distribution,

$$\left[\widehat{R}_N(\tilde{\theta}) - Q_c\frac{\widehat{\text{std}}_N(\tilde{\theta})}{\sqrt{N}}, \widehat{R}_N(\tilde{\theta}) + Q_c\frac{\widehat{\text{std}}_N(\tilde{\theta})}{\sqrt{N}}\right]$$

is a confidence interval for $\mathbb{E}\left[h^2(s, f_{\tilde{\theta}})\right]$ with asymptotic confidence level c. We also introduce

$$\widehat{\mathcal{R}}_{N,\text{rel}}(\tilde{\theta}) = \frac{\widehat{R}_N(\hat{\theta})}{\widehat{R}_N(\tilde{\theta})} - 1$$

in order to make the comparison of our estimator $\hat{\theta}$ and the estimator $\tilde{\theta}$ easier. When $\mathcal{R}_{\text{rel}}(\tilde{\theta})$ is negative our estimator is better than $\tilde{\theta}$ whereas if $\mathcal{R}_{\text{rel}}(\tilde{\theta})$ is positive, our estimator is worse than $\tilde{\theta}$. More precisely, if $\mathcal{R}_{\text{rel}}(\tilde{\theta}) = \alpha$, the risk of our estimator corresponds to the one of $\tilde{\theta}$ reduced of $100|\alpha|\%$ when $\alpha < 0$ and increased of $100\alpha\%$ when $\alpha > 0$.

The results are gathered below.

		$n = 10$	$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	$\widehat{R}_{10^6}(\hat{\theta})$	0.0130	0.0051	0.0025	0.0017	0.0013
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0129	0.0051	0.0025	0.0017	0.0013
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mle}})$	$6 \cdot 10^{-4}$	10^{-5}	$7 \cdot 10^{-7}$	$-8 \cdot 10^{-9}$	$2 \cdot 10^{-9}$
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0192	0.0073	0.0036	0.0024	0.0018
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0192	0.0073	0.0036	0.0024	0.0018
Example 2	$\widehat{R}_{10^6}(\hat{\theta})$	0.0123	0.0050	0.0025	0.0017	0.0012
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0123	0.0050	0.0025	0.0017	0.0012
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mle}})$	$5 \cdot 10^{-10}$	$9 \cdot 10^{-10}$	$-2 \cdot 10^{-9}$	$-2 \cdot 10^{-9}$	$-3 \cdot 10^{-9}$
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0170	0.0070	0.0035	0.0023	0.0018
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0170	0.0070	0.0035	0.0023	0.0018
Example 3	$\widehat{R}_{10^6}(\hat{\theta})$	0.0130	0.0051	0.0025	0.0017	0.0013
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0129	0.0051	0.0025	0.0017	0.0013
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mle}})$	$6 \cdot 10^{-4}$	$2 \cdot 10^{-5}$	10^{-6}	-10^{-7}	$-4 \cdot 10^{-9}$
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0192	0.0073	0.0036	0.0024	0.0018
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0192	0.0073	0.0036	0.0024	0.0018
Example 4	$\widehat{R}_{10^6}(\hat{\theta})$	0.0152	0.0054	0.0026	0.0017	0.0013
	$\widehat{R}_{10^4}(\tilde{\theta}_{\text{mle}})$	0.0149	0.0054	0.0026	0.0017	0.0012
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\theta}_{\text{mle}})$	-0.001	$-2 \cdot 10^{-4}$	-10^{-8}	$-3 \cdot 10^{-8}$	$9 \cdot 10^{-8}$
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0267	0.0083	0.0038	0.0025	0.0018
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0255	0.0083	0.0039	0.0025	0.0018

Example 5	$\widehat{R}_{10^6}(\hat{\theta})$	0.0468	0.0192	0.0096	0.0064	0.0048
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0476	0.0196	0.0099	0.0066	0.0050
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mvub}})$	0.0350	0.0144	0.0073	0.0049	0.0037
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mle}})$	-0.0160	-0.0202	-0.0287	-0.0271	-0.0336
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mvub}})$	0.3390	0.3329	0.3215	0.3243	0.3148
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0529	0.0223	0.0112	0.0075	0.0056
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0453	0.0192	0.0098	0.0066	0.0049
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mvub}})$	0.0316	0.0132	0.0067	0.0045	0.0034
Example 6	$\widehat{R}_{10^6}(\hat{\theta})$	0.0504	0.0197	0.0098	0.0065	0.0049
	$\widehat{R}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0483	0.0197	0.0099	0.0066	0.0050
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}_{\text{mle}})$	0.0436	-0.0019	-0.0180	-0.0242	-0.0263
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0597	0.0233	0.0115	0.0076	0.0057
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}_{\text{mle}})$	0.0467	0.0195	0.0099	0.0066	0.0050
Example 7	$\widehat{R}_{10^6}(\hat{\theta})$	0.0455	0.0193	0.0098	0.0066	0.0050
	$\widehat{R}_{10^6}(\tilde{\theta}')$	0.0454	0.0192	0.0098	0.0066	0.0050
	$\widehat{\mathcal{R}}_{10^6,\text{rel}}(\tilde{\theta}')$	0.0029	0.0029	0.0031	0.0028	0.0030
	$\widehat{\text{std}}_{10^6}(\hat{\theta})$	0.0416	0.0186	0.0096	0.0065	0.0049
	$\widehat{\text{std}}_{10^6}(\tilde{\theta}')$	0.0415	0.0185	0.0096	0.0065	0.0049
Example 8	$\widehat{R}_{10^4}(\hat{\theta})$	0.050	0.022	0.012	0.008	0.006
	$\widehat{R}_{10^4}(\tilde{\theta}_{\text{mean}})$	0.084	0.061	0.049	0.043	0.039
	$\widehat{R}_{10^4}(\tilde{\theta}_{\text{median}})$	0.066	0.036	0.025	0.019	0.017
	$\widehat{R}_{10^4}(\tilde{\theta}_{\text{mspe}})$	0.050	0.022	0.012	0.008	0.006
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\theta}_{\text{mean}})$	-0.40	-0.64	-0.76	-0.82	-0.85
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\theta}_{\text{median}})$	-0.25	-0.39	-0.54	-0.59	-0.65
	$\widehat{\text{std}}_{10^4}(\hat{\theta})$	0.054	0.025	0.013	0.009	0.007
	$\widehat{\text{std}}_{10^4}(\tilde{\theta}_{\text{mean}})$	0.045	0.032	0.025	0.022	0.020
	$\widehat{\text{std}}_{10^4}(\tilde{\theta}_{\text{median}})$	0.052	0.032	0.020	0.016	0.014
	$\widehat{\text{std}}_{10^4}(\tilde{\theta}_{\text{mspe}})$	0.051	0.025	0.014	0.009	0.007

In the first four examples, the risk of our estimator is very close to the maximum likelihood estimator one, whatever the value of n. In Example 5, our estimator slightly improves the maximum likelihood estimator but is worse than the minimum variance unbiased estimator. In Example 6, the risk of our estimator is larger than the one of the m.l.e when $n = 10$ but is slightly smaller as soon as n becomes larger than 25. In Example 7, the risk of our estimator is 0.3% larger than the one of $\tilde{\theta}'$. In Example 8, our estimator significantly improves the empirical mean and the median. Its risk is comparable to the one of the m.s.p.e (we omit in this example the value of $\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\theta}_{\text{mspe}})$ because it is influenced by the procedure we have used to build the m.s.p.e).

When the model is regular enough, these simulations show that our estimation strategy provides an estimator whose risk is almost equal to the one of the maximum likelihood estimator.

Moreover, our estimator seems to work rather well in a model where the m.l.e does not exist (case of Example 8). Remark that contrary to the maximum likelihood method, our procedure does not involve the search of a global maximum.

We now bring to light the connection between our estimator and the m.l.e when the model is regular enough (that is in the first four examples). Let for $c \in \{0.99, 0.999, 1\}$, q_c be the c-quantile of the random variable $\left|\hat{\theta} - \tilde{\theta}_{\text{mle}}\right|$, and $\hat{q}_c$ be the empirical version based on N samples ($N = 10^6$ in examples 1,2,3 and $N = 10^4$ in Example 4). The results are the following.

		$n = 10$	$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	$\hat{q}_{0.99}$	10^{-7}	10^{-7}	10^{-7}	10^{-7}	10^{-7}
	$\hat{q}_{0.999}$	0.07	10^{-7}	10^{-7}	10^{-7}	10^{-7}
	$\hat{q}_1$	1.9	0.3	0.06	0.005	10^{-7}
Example 2	$\hat{q}_{0.99}$	$2 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$
	$\hat{q}_{0.999}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$
	$\hat{q}_1$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-7}$
Example 3	$\hat{q}_{0.99}$	10^{-7}	10^{-7}	10^{-7}	10^{-7}	10^{-7}
	$\hat{q}_{0.999}$	0.03	10^{-7}	10^{-7}	10^{-7}	10^{-7}
	$\hat{q}_1$	0.38	0.12	0.01	0.007	10^{-7}
Example 4	$\hat{q}_{0.99}$	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
	$\hat{q}_{0.999}$	$3 \cdot 10^{-6}$	10^{-6}	10^{-6}	10^{-6}	10^{-6}
	$\hat{q}_1$	1.5	0.1	10^{-6}	10^{-6}	10^{-6}

This array shows that with large probability our estimator is very close to the m.l.e. This probability is quite high for small values of n and even more for larger values of n. This explains why the risks of these two estimators are very close in the first four examples. Note that the value of η prevents the empirical quantile from being lower than something of order 10^{-7} according to the examples (in Example 4, the value of 10^{-6} is due to the way we have built the m.l.e).

4.4. Speed of the procedure. For the sake of completeness, we specify below the number of tests that have been calculated in the preceding examples.

	$n = 10$	$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	77 (1.4)	77 (0.9)	77 (0.7)	77 (0.6)	77 (0.5)
Example 2	293 (1)	294 (1)	294 (0.9)	295 (0.9)	295 (0.9)
Example 3	89 (0.75)	90 (0.5)	90 (0.5)	90 (0.5)	90 (0.5)
Example 4	100 (3.5)	100 (0.5)	100 (0.001)	100 (0)	100 (0)
Example 5	460 (3)	461 (1)	462 (0.6)	462 (0.4)	462 (0.3)
Example 6	687 (0)	687 (0)	687 (0)	687 (0)	687 (0)
Example 7	412 (8)	419 (8)	425 (8)	429 (8)	432 (8)
Example 8	173209 (10)	173212 (0)	173212 (0.9)	173206 (12)	173212 (0.3)

Figure 4.1: Number of tests computed averaged over 10^6 samples for examples 1 to 7 and over 10^4 samples for example 8. The corresponding standard deviations are in brackets.

4.5. Simulations when $s \notin \mathscr{F}$. In Section 4.3, we were in the favourable situation where the true distribution s belonged to the model $\mathscr{F}$, which may not hold true in practice. We now work with random variables $X_1, \dots, X_n$ simulated according to a density $s \notin \mathscr{F}$ to illustrate the robustness properties of our estimator.

We begin with an example proposed in Birgé (2006). We generate $X_1, \dots, X_n$ according to the density

$$s(x) = 10\left[(1-2n^{-1})\mathbb{1}_{[0,1/10]}(x) + 2n^{-1}\mathbb{1}_{[9/10,1]}(x)\right] \quad \text{for all } x \in \mathbb{R}$$

and compare our estimator to the maximum likelihood estimator for the uniform model

$$\mathscr{F} = \{f_\theta,\, \theta \in [0.01, 10]\} \quad \text{where} \quad f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}. \tag{13}$$

It is worthwhile to notice that $h^2(s, \mathscr{F}) = \mathcal{O}(n^{-1})$, which means that s is close to $\mathscr{F}$ when n is large, and that our estimator still satisfies $\mathbb{E}[h^2(s, f_{\hat{\theta}})] = \mathcal{O}(n^{-1})$. Contrary to our estimator, the outliers make the m.l.e unstable as shown in the array below.

	$n=10$	$n=25$	$n=50$	$n=75$	$n=100$
$\widehat{R}_N(\hat{\theta})$	0.20	0.06	0.03	0.02	0.015
$\widehat{R}_N(\tilde{\theta}_{\text{mle}})$	0.57	0.56	0.56	0.56	0.57

Figure 4.2: Risks for simulated data averaged over 10^4 samples.

We now propose a second example based on the mixture of two uniform laws. We use the same statistical model $\mathscr{F}$ but we modify the distribution of the observations. We take $p \in (0,1)$ and define the true underlying density by

$$s_p(x) = (1-p)f_1(x) + pf_2(x) \quad \text{for all } x \in \mathbb{R}.$$

Set $p_0 = 1 - 1/\sqrt{2}$. One can check that

$$\begin{aligned} H^2(s_p, \mathscr{F}) &= \begin{cases} H^2(s_p, f_1) & \text{if } p \le p_0 \\ H^2(s_p, f_2) & \text{if } p > p_0, \end{cases} \\ &= \begin{cases} 1 - \sqrt{2-p}/\sqrt{2} & \text{if } p \le p_0 \\ 1 - (\sqrt{2-p} + \sqrt{p})/2 & \text{if } p > p_0, \end{cases} \end{aligned}$$

which means that the best estimator of $\mathscr{F}$ is f_1 when $p < p_0$ and f_2 when $p > p_0$.

We now compare our estimator $\hat{\theta}$ to the m.l.e $\tilde{\theta}_{\text{mle}}$. For a lot of values of p, we simulate N samples of n random variables with density s_p and investigate the behaviour of the estimator $\tilde{\theta} \in \{\hat{\theta}, \tilde{\theta}_{\text{mle}}\}$ by computing the function

$$\widehat{R}_{p,n,N}(\tilde{\theta}) = \frac{1}{N}\sum_{i=1}^{N} h^2(s_p, f_{\tilde{\theta}^{(p,i)}})$$

where $\tilde{\theta}^{(p,i)}$ is the value of the estimator $\tilde{\theta}$ corresponding to the i^{th} sample whose density is s_p. We draw below the functions $p \mapsto \widehat{R}_{p,n,N}(\hat{\theta})$, $p \mapsto \widehat{R}_{p,n,N}(\tilde{\theta}_{\text{mle}})$ and $p \mapsto H^2(s_p, \mathscr{F})$ for $n = 100$ and then for $n = 10^4$.

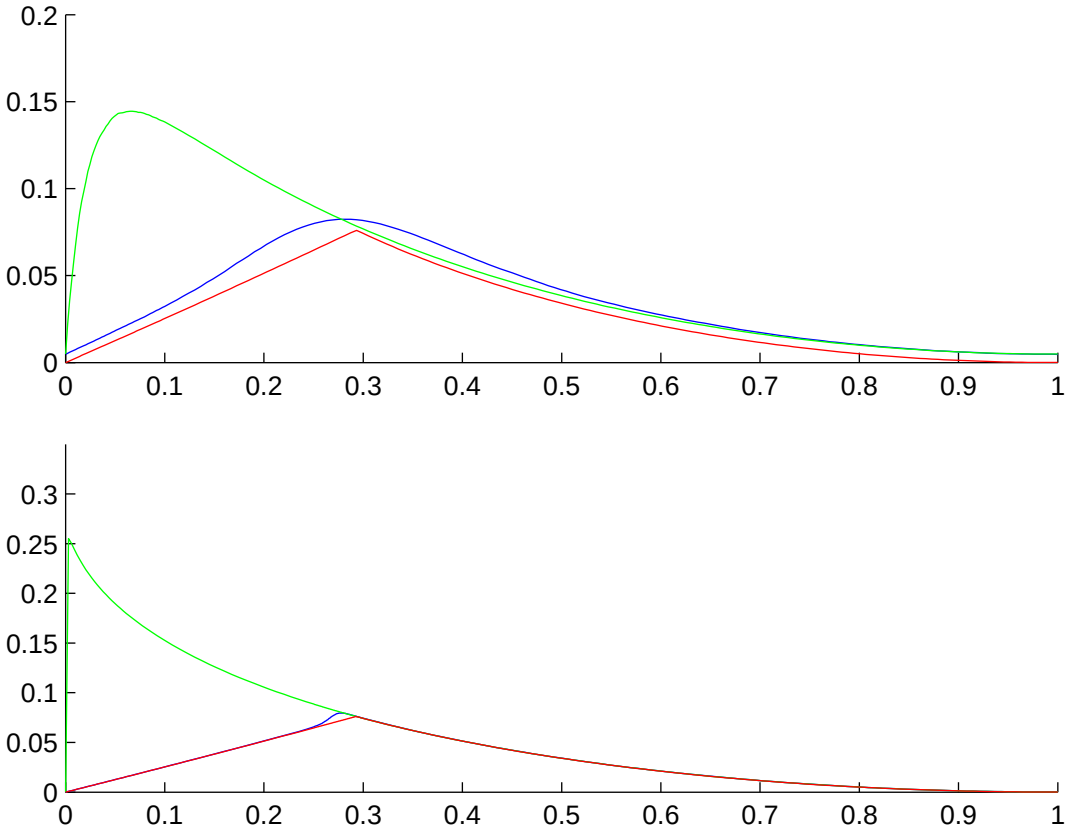


Figure 4.3: Red: $p \mapsto H^2(s_p, \mathscr{F})$. Blue: $p \mapsto \widehat{R}_{p,n,5000}(\hat{\theta})$. Green: $p \mapsto \widehat{R}_{p,n,5000}(\tilde{\theta}_{\text{mle}})$.

We observe that the m.l.e is rather good when $p \geq p_0$ and very poor when $p < p_0$. This can be explained by the fact that the m.l.e $\tilde{\theta}_{\text{mle}}$ is close to 2 as soon as the number n of observations is large enough. The shape of the function $p \mapsto \widehat{R}_{p,n,5000}(\hat{\theta})$ is quite more satisfying since it looks more like the function $p \mapsto H^2(s_p, \mathscr{F})$. The lower figure suggests that $\widehat{R}_{p,n,N}(\hat{\theta})$ converges to $H^2(s_p, \mathscr{F})$ when n, N go to infinity except on a small neighbourhood before p_0.

5. Models parametrized by a multidimensional parameter

5.1. Assumption. We now deal with models $\mathscr{F} = \{f_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ indexed by a rectangle

$$\Theta = \prod_{j=1}^{d} [m_j, M_j]$$

of $\mathbb{R}^d$ with d larger than 2. Assumption 1 is supposed to be fulfilled all along this section.

5.2. Definition of the test. As previously, our estimation strategy is based on the existence for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ of a measurable function $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ of the observations possessing suitable sta-

tistical properties. The definition of this functional is the natural extension of the one we have proposed in Section 3.2.

Let for $j \in \{1, \dots, d\}$, $t_j \in (0, d^{1/\alpha_j}]$ and $\epsilon_j = t_j(\overline{R}n)^{-1/\alpha_j}$. We then define the finite sets

$$\begin{aligned} \Theta_{\text{dis}} &= \left\{ (m_1 + k_1\epsilon_1, \dots, m_d + k_d\epsilon_d)\,,\ \forall j \in \{1, \dots, d\},\ k_j \leq (M_j - m_j)\epsilon_j^{-1} \right\} \\ \mathscr{F}_{\text{dis}} &= \{f_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta_{\text{dis}}\}. \end{aligned}$$

Let π be the map defined on $\prod_{j=1}^d [m_j, M_j]$ by

$$\pi(\boldsymbol{x}) = \big(m_1 + \lfloor (x_1 - m_1)/\epsilon_1 \rfloor \epsilon_1, \dots, m_d + \lfloor (x_d - m_d)/\epsilon_d \rfloor \epsilon_d\big) \quad \text{for all } \boldsymbol{x} = (x_1, \dots, x_d) \in \prod_{j=1}^d [m_j, M_j]$$

where $\lfloor \cdot \rfloor$ is the integer part. We then define $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ between two elements $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ by

$$T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \overline{T}(f_{\pi(\boldsymbol{\theta})}, f_{\pi(\boldsymbol{\theta}')}) \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta = \prod_{j=1}^d [m_j, M_j] \tag{14}$$

where $\overline{T}$ is the functional given by (4).

5.3. Basic ideas. For the sake of simplicity, we first consider the case $d = 2$. We shall build a decreasing sequence $(\Theta_i)_i$ of rectangles by induction. When there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $s = f_{\boldsymbol{\theta}_0}$, these rectangles Θ_i can be interpreted as confidence sets for $\boldsymbol{\theta}_0$. Their construction is strongly inspired from the heuristics of Section 3.1.

We set $\Theta_1 = \Theta$. Assume that $\Theta_i = [a_1, b_1] \times [a_2, b_2]$ and let us explain how we can build a confidence set $\Theta_{i+1} = [a_1, b_1] \times [a_2', b_2']$ with a_2', b_2' satisfying $b_2' - a_2' < b_2 - a_2$.

We begin to build by induction two preliminary finite sequences $(\boldsymbol{\theta}^{(j)})_{1 \leq j \leq N}$, $(\boldsymbol{\theta}'^{(j)})_{1 \leq j \leq N}$ of elements of $\mathbb{R}^2$. Let $\boldsymbol{\theta}^{(1)} = (a_1, b_1)$ be the bottom left-hand corner of Θ_i and $\boldsymbol{\theta}'^{(1)} = (a_1, b_2)$ be the top left-hand corner of Θ_i. Let $\bar{r}_1(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)})$, $\bar{r}_2(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)})$, $\bar{r}_1(\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}^{(1)})$, $\underline{r}_2(\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}^{(1)})$ be positive numbers such that the rectangles

$$\begin{aligned} \mathcal{R}_1 &= [a_1, a_1 + \bar{r}_1(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)})] \times [a_2, a_2 + \bar{r}_2(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)})] \\ \mathcal{R}_1' &= [a_1, a_1 + \bar{r}_1(\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}^{(1)})] \times [b_2 - \underline{r}_2(\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}^{(1)}), b_2] \end{aligned}$$

are respectively included in the Hellinger balls

$$\mathcal{B}\big(\boldsymbol{\theta}^{(1)}, \kappa^{1/2} h(f_{\boldsymbol{\theta}^{(1)}}, f_{\boldsymbol{\theta}'^{(1)}})\big) \quad \text{and} \quad \mathcal{B}\big(\boldsymbol{\theta}'^{(1)}, \kappa^{1/2} h(f_{\boldsymbol{\theta}^{(1)}}, f_{\boldsymbol{\theta}'^{(1)}})\big).$$

See (3) for the precise definition of these balls.

We define $\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)} \in \mathbb{R}^2$ as follows

$$\begin{aligned} \boldsymbol{\theta}^{(2)} &= \begin{cases} \boldsymbol{\theta}^{(1)} + (\bar{r}_1(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)}), 0) & \text{if } T(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)}) \geq 0 \\ \boldsymbol{\theta}^{(1)} & \text{otherwise} \end{cases} \\ \boldsymbol{\theta}'^{(2)} &= \begin{cases} \boldsymbol{\theta}'^{(1)} + (\bar{r}_1(\boldsymbol{\theta}'^{(1)}, \boldsymbol{\theta}^{(1)}), 0) & \text{if } T(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)}) \leq 0 \\ \boldsymbol{\theta}'^{(1)} & \text{otherwise.} \end{cases} \end{aligned}$$

Here is an illustration.

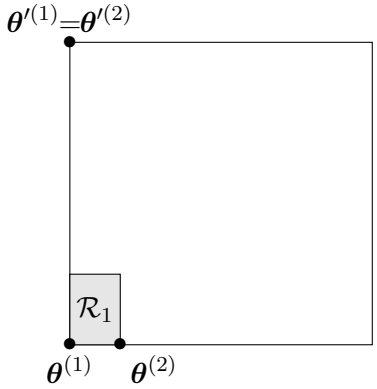


Figure 4.4: Construction of $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\theta}'^{(2)}$ when $T(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}'^{(1)}) > 0$.

It is worthwhile to notice that in this figure, the heuristics of Section 3.1 suggest that $\boldsymbol{\theta}_0$ belongs to $\Theta_i \setminus \mathcal{R}_1$.

Now, if the first component of $\boldsymbol{\theta}^{(2)} = (\theta_1^{(2)}, \theta_2^{(2)})$ is larger than b_1, that is $\theta_1^{(2)} \geq b_1$, we set $N = 1$ and stop the construction of the vectors $\boldsymbol{\theta}^{(i)}$, $\boldsymbol{\theta}'^{(i)}$. Similarly, if $\theta_1'^{(2)} \geq b_1$, we set $N = 1$ and stop the construction of the $\boldsymbol{\theta}^{(i)}$, $\boldsymbol{\theta}'^{(i)}$.

If $\theta_1^{(2)} < b_1$ and $\theta_1'^{(2)} < b_1$, we consider positive numbers $\bar{r}_1(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)})$, $\bar{r}_2(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)})$, $\bar{r}_1(\boldsymbol{\theta}'^{(2)}, \boldsymbol{\theta}^{(2)})$, $\underline{r}_2(\boldsymbol{\theta}'^{(2)}, \boldsymbol{\theta}^{(2)})$ such that the rectangles

$$\begin{aligned} \mathcal{R}_2 &= [\theta_1^{(2)}, \theta_1^{(2)} + \bar{r}_1(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)})] \times [a_2, a_2 + \bar{r}_2(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)})] \\ \mathcal{R}_2' &= [\theta_1'^{(2)}, \theta_1'^{(2)} + \bar{r}_1(\boldsymbol{\theta}'^{(2)}, \boldsymbol{\theta}^{(2)})] \times [b_2 - \underline{r}_2(\boldsymbol{\theta}'^{(2)}, \boldsymbol{\theta}^{(2)}), b_2] \end{aligned}$$

are respectively included in the Hellinger balls

$$\mathcal{B}\big(\boldsymbol{\theta}^{(2)}, \kappa^{1/2} h(f_{\boldsymbol{\theta}^{(2)}}, f_{\boldsymbol{\theta}'^{(2)}})\big) \quad \text{and} \quad \mathcal{B}\big(\boldsymbol{\theta}'^{(2)}, \kappa^{1/2} h(f_{\boldsymbol{\theta}^{(2)}}, f_{\boldsymbol{\theta}'^{(2)}})\big).$$

We then define $\boldsymbol{\theta}^{(3)}, \boldsymbol{\theta}'^{(3)} \in \mathbb{R}^2$ by

$$\begin{aligned} \boldsymbol{\theta}^{(3)} &= \begin{cases} \boldsymbol{\theta}^{(2)} + (\bar{r}_1(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)}), 0) & \text{if } T(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)}) \geq 0 \\ \boldsymbol{\theta}^{(2)} & \text{otherwise} \end{cases} \\ \boldsymbol{\theta}'^{(3)} &= \begin{cases} \boldsymbol{\theta}'^{(2)} + (\bar{r}_1(\boldsymbol{\theta}'^{(2)}, \boldsymbol{\theta}^{(2)}), 0) & \text{if } T(\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}'^{(2)}) \leq 0 \\ \boldsymbol{\theta}'^{(2)} & \text{otherwise.} \end{cases} \end{aligned}$$

If $\theta_1^{(3)} \geq b_1$ or if $\theta_1'^{(3)} \geq b_1$ we stop the construction and set $N = 2$. In the contrary case, we repeat this step to build the vectors $\boldsymbol{\theta}^{(4)}$ and $\boldsymbol{\theta}'^{(4)}$.

We repeat these steps until the construction stops. Let N be the integer for which $\theta_1^{(N+1)} \geq b_1$

or $\theta_1'^{(N+1)} \geq b_1$. We then define

$$a_2' = \begin{cases} a_2 + \min_{1\leq j\leq N} \bar{r}_2(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}'^{(j)}) & \text{if } \theta_1^{(N+1)} \geq b_1 \\ a_2 & \text{otherwise} \end{cases}$$

$$b_2' = \begin{cases} b_2 - \min_{1\leq j\leq N} \underline{r}_2(\boldsymbol{\theta}'^{(j)}, \boldsymbol{\theta}^{(j)}) & \text{if } \theta_1'^{(N+1)} \geq b_1 \\ b_2 & \text{otherwise} \end{cases}$$

and set $\Theta_{i+1} = [a_1, b_1] \times [a_2', b_2']$.

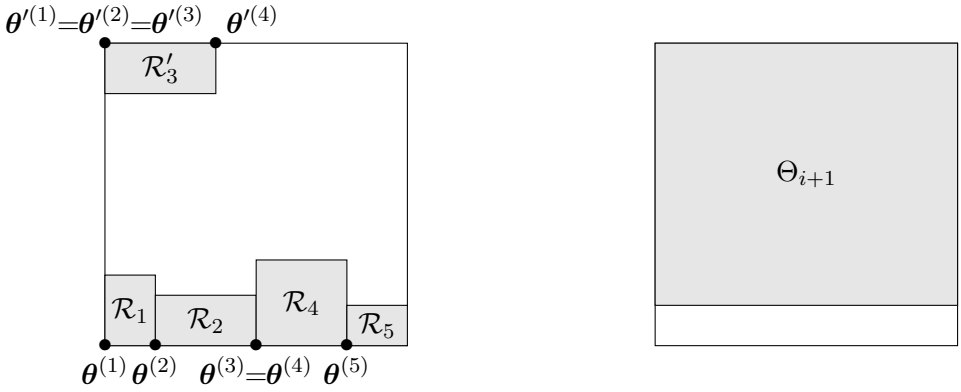


Figure 4.5: Illustration when $N = 5$, $T(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}'^{(i)}) > 0$ for $i \in \{1, 2, 4, 5\}$ and $T(\boldsymbol{\theta}^{(3)}, \boldsymbol{\theta}'^{(3)}) < 0$.

In this figure, the set

$$\Theta_i \setminus \big(\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_4 \cup \mathcal{R}_5\big)$$

is a confidence set for $\boldsymbol{\theta}_0$. The set Θ_{i+1} is the smallest rectangle containing this confidence set.

Remark 1. We define Θ_{i+1} as a rectangle to make the procedure easier to implement.

Remark 2. By using a similar strategy, we can also build a confidence set Θ_{i+1} of the form $\Theta_{i+1} = [a_1', b_1'] \times [a_2, b_2]$ where a_1', b_1' are such that $b_1' - a_1' < b_1 - a_1$.

We shall build the rectangles Θ_i until their diameters become sufficiently small. The estimator we shall consider will be the center of the last rectangle built.

5.4. Procedure. In the general case, that is when $d \geq 2$, we build a finite sequence of rectangles $(\Theta_i)_i$ of $\Theta = \prod_{j=1}^d [m_j, M_j]$. We consider $\kappa > 0$ and for all rectangle $\mathcal{C} = \prod_{j=1}^d [a_j, b_j] \subset \Theta$, vectors $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}$, and integers $j \in \{1, \ldots, d\}$, we introduce positive numbers $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ such that

$$\mathcal{C} \bigcap \prod_{j=1}^d \big[\theta_j - \underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}'), \theta_j + \bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')\big] \subset \mathcal{B}\big(\boldsymbol{\theta}, \kappa^{1/2} h(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})\big). \tag{15}$$

We also consider for all $j \in \{1, \dots, d\}$, $\underline{R}_{\mathcal{C},j} \geq \underline{R}_j$ such that

$$h^2\left(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}\right) \geq \sup_{1 \leq j \leq d} \underline{R}_{\mathcal{C},j} |\theta_j - \theta'_j|^{\alpha_j} \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}. \tag{16}$$

We finally consider for all $j \in \{1, \dots, d\}$, an one-to-one map ψ_j from $\{1, \dots, d-1\}$ into $\{1, \dots, d\} \setminus \{j\}$.

We set $\Theta_1 = \Theta$. Given Θ_i, we define Θ_{i+1} by using the following algorithm.

Algorithm 2 Definition of Θ_{i+1} from Θ_i

Require: $\Theta_i = \prod_{j=1}^{d} [a_j, b_j]$

Choose $k \in \{1, \dots, d\}$ such that

$$\underline{R}_{\Theta_i,k}(b_k - a_k)^{\alpha_k} = \max_{1 \leq j \leq d} \underline{R}_{\Theta_i,j}(b_j - a_j)^{\alpha_j}$$

$\boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \leftarrow (a_1, \dots, a_d)$, $\boldsymbol{\theta}' = (\theta'_1, \dots, \theta'_d) \leftarrow \boldsymbol{\theta}$ and $\theta'_k \leftarrow b_k$

$\varepsilon_j \leftarrow \bar{r}_{\Theta_i,j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\varepsilon'_j \leftarrow \bar{r}_{\Theta_i,j}(\boldsymbol{\theta}', \boldsymbol{\theta})$ for all $j \neq k$

$\varepsilon_k \leftarrow (b_k - a_k)/2$ and $\varepsilon'_k \leftarrow (b_k - a_k)/2$

repeat

  Test $\leftarrow T(\boldsymbol{\theta}, \boldsymbol{\theta}')$

  For all j, $\bar{r}_j \leftarrow \bar{r}_{\Theta_i,j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\bar{r}'_j \leftarrow \bar{r}_{\Theta_i,j}(\boldsymbol{\theta}', \boldsymbol{\theta})$, $\underline{r}'_j \leftarrow \underline{r}_{\Theta_i,j}(\boldsymbol{\theta}', \boldsymbol{\theta})$

  if Test ≥ 0 **then**

    $\varepsilon_{\psi_k(1)} \leftarrow \bar{r}_{\psi_k(1)}$

    $\varepsilon_{\psi_k(j)} \leftarrow \min(\varepsilon_{\psi_k(j)}, \bar{r}_{\psi_k(j)})$ for all $j \in \{2, \dots, d-1\}$

    $\varepsilon_k \leftarrow \min(\varepsilon_k, \bar{r}_k)$

    $J \leftarrow \left\{1 \leq j \leq d-1,\ \theta_{\psi_k(j)} + \varepsilon_{\psi_k(j)} < b_{\psi_k(j)}\right\}$

    if $J \neq \emptyset$ **then**

      $\mathrm{j}_{\min} \leftarrow \min J$

      $\theta_{\psi_k(j)} \leftarrow a_{\psi_k(j)}$ for all $j \leq \mathrm{j}_{\min} - 1$

      $\theta_{\psi_k(\mathrm{j}_{\min})} \leftarrow \theta_{\psi_k(\mathrm{j}_{\min})} + \varepsilon_{\psi_k(\mathrm{j}_{\min})}$

    else

      $\mathrm{j}_{\min} \leftarrow d$

    end if

  end if

  if Test ≤ 0 **then**

    $\varepsilon'_{\psi_k(1)} \leftarrow \bar{r}'_{\psi_k(1)}$

    $\varepsilon'_{\psi_k(j)} \leftarrow \min(\varepsilon'_{\psi_k(j)}, \bar{r}'_{\psi_k(j)})$ for all $j \in \{2, \dots, d-1\}$

    $\varepsilon'_k \leftarrow \min(\varepsilon'_k, \underline{r}'_k)$

    $J' \leftarrow \left\{1 \leq j' \leq d-1,\ \theta'_{\psi_k(j')} + \varepsilon'_{\psi_k(j')} < b_{\psi_k(j')}\right\}$

    if $J' \neq \emptyset$ **then**

      $\mathrm{j}'_{\min} \leftarrow \min J'$

      $\theta'_{\psi_k(j)} \leftarrow a_{\psi_k(j)}$ for all $j \leq \mathrm{j}'_{\min} - 1$

      $\theta'_{\psi_k(\mathrm{j}'_{\min})} \leftarrow \theta'_{\psi_k(\mathrm{j}'_{\min})} + \varepsilon'_{\psi_k(\mathrm{j}'_{\min})}$

    else

      $\mathrm{j}'_{\min} \leftarrow d$

    end if

33: **end if**
34: **until** $\mathrm{j}_{\min} = d$ or $\mathrm{j}'_{\min} = d$
35: **if** $\mathrm{j}_{\min} = d$ **then**
36: $a_k \leftarrow a_k + \varepsilon_k$
37: **end if**
38: **if** $\mathrm{j}'_{\min} = d$ **then**
39: $b_k \leftarrow b_k - \varepsilon'_k$
40: **end if**
41: $\Theta_{i+1} \leftarrow \prod_{j=1}^{d} [a_j, b_j]$
42: **Return:** Θ_{i+1}

We now consider d positive numbers $\eta_1, \dots, \eta_d$ and use the algorithm below to build our estimator $\hat{\boldsymbol{\theta}}$.

Algorithm 3

1: Set $a_j = m_j$ and $b_j = M_j$ for all $j \in \{1, \dots, d\}$
2: $i \leftarrow 0$
3: **while** There exists $j \in \{1, \dots, d\}$ such that $b_j - a_j > \eta_j$ **do**
4: $i \leftarrow i + 1$
5: Build Θ_i and set $a_1, \dots, a_d$, $b_1, \dots, b_d$ such that $\prod_{j=1}^{d} [a_j, b_j] = \Theta_i$
6: **end while**
7: **Return:** $\hat{\boldsymbol{\theta}} = ((a_1 + b_1)/2, \dots, (a_d + b_d)/2)$

The parameters κ, t_j, η_j $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be interpreted as in dimension 1. We have introduced a new parameter $\underline{R}_{\mathcal{C},j}$ whose role is to control more accurately the Hellinger distance in order to make the procedure faster. Sometimes, the computation of this parameter is difficult in practice, in which case we can avoid it by proceeding as follows. For all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,

$$h^2 \left(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'} \right) \geq \sup_{1 \leq j \leq d} \underline{R} |\theta_j - \theta'_j|^{\alpha_j}$$

where $\underline{R} = \min_{1 \leq j \leq d} \underline{R}_j$, which means that we can always assume that $\underline{R}_j$ is independent of j. Choosing $\underline{R}_{\Theta_i, j} = \underline{R}$ simplifies the only line where this parameter is involved (line 1 of Algorithm 2). It becomes

$$(b_k - a_k)^{\alpha_k} = \max_{1 \leq j \leq d} (b_j - a_j)^{\alpha_j}$$

and k can be calculated without computing $\underline{R}$.

5.5. Risk bound. Suitable values of the parameters lead to a risk bound for our estimator $\hat{\boldsymbol{\theta}}$.

Theorem 5. *Suppose that Assumption 1 holds. Let $\bar{\kappa}$ be defined by (7), and assume that $\kappa \in (0, \bar{\kappa})$, and for all $j \in \{1, \dots, d\}$, $t_j \in (0, d^{1/\alpha_j}]$,*

$$\epsilon_j = t_j (\overline{R}_j n)^{-1/\alpha_j}, \quad \eta_j \in [\epsilon_j, d^{1/\alpha_j} (\overline{R}_j n)^{-1/\alpha_j}].$$

Suppose that for all rectangle $\mathcal{C}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}$, the numbers $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, are such that (15) holds.

Then, for all $\xi > 0$, the estimator $\hat{\boldsymbol{\theta}}$ built by Algorithm 3 satisfies

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\boldsymbol{\theta}}}) \geq h^2(s, \mathscr{F}) + \frac{d}{n} + \xi\right] \leq e^{-n\xi}$$

where $C > 0$ depends only on κ, $(\overline{R}_j/\underline{R}_j)_{1\leq j\leq d}$, $(\alpha_j)_{1\leq j\leq d}$, $(t_j)_{1\leq j\leq d}$.

Remark. A look at the proof of the theorem shows that Theorem 1 ensues from this theorem when $t_j = d^{1/\alpha_j}$ and $\eta_j = \epsilon_j$.

5.6. Choice of $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$. The parameters $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ are involved in the procedure and must be calculated. They may be chosen arbitrary provided that the rectangle

$$\mathcal{C} \bigcap \prod_{j=1}^{d} \left[\theta_j - \underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}'), \theta_j + \bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')\right]$$

is included in the Hellinger ball $\mathcal{B}\big(\boldsymbol{\theta}, \kappa^{1/2} h(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})\big)$. Indeed, the theoretical properties of the estimator given by the preceding theorem does not depend on these values.

However, the numerical complexity of the algorithm strongly depends on these parameters. The algorithm computes less tests when $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ are large and we have thus an interest in defining them as the largest numbers possible. In the cases where a direct computation of these numbers is difficult, we may use a similar strategy that the one adopted in the unidimensional case (Section 3.5).

First way. We may consider $(\overline{R}_{\mathcal{C},1}, \ldots, \overline{R}_{\mathcal{C},d}) \in \prod_{j=1}^{d} (0, \overline{R}_j]$ such that

$$h^2\left(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}\right) \leq \sup_{1\leq j\leq d} \overline{R}_{\mathcal{C},j} |\theta_j - \theta'_j|^{\alpha_j} \quad \text{for all } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C} \tag{17}$$

and define them by

$$\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left((\kappa/\overline{R}_{\mathcal{C},j}) h^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})\right)^{1/\alpha_j}. \tag{18}$$

One can verify that this definition implies (15).

Second way. An alternative definition that does not involve the Hellinger distance is

$$\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left(\kappa/\overline{R}_{\mathcal{C},j} \sup_{1\leq k\leq d} \underline{R}_{\mathcal{C},k} |\theta'_k - \theta_k|^{\alpha_k}\right)^{1/\alpha_j}. \tag{19}$$

Similarly, one can check that (15) holds.

The complexity of our procedure can be upper-bounded as soon as $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ are large enough.

Proposition 6. *Suppose that the assumptions of Theorem 5 are fulfilled and that for all $j \in \{1,\dots,d\}$, all rectangle $\mathcal{C}$, $\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathcal{C}$, the numbers $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$ are larger than*

$$\left(\kappa/\overline{R}_{\mathcal{C},j} \sup_{1\le k\le d} \underline{R}_{\mathcal{C},k}|\theta'_k-\theta_k|^{\alpha_k}\right)^{1/\alpha_j} \tag{20}$$

where the $\underline{R}_{\mathcal{C},j}$ and $\overline{R}_{\mathcal{C},j}$ are respectively such that (16) and (17) hold and such that $\underline{R}_{\mathcal{C},j} \ge \underline{R}_j$ and $\overline{R}_{\mathcal{C},j} \le \overline{R}_j$.

Then, the number of tests computed to build the estimator $\hat{\boldsymbol{\theta}}$ is smaller than

$$4\left[\prod_{j=1}^{d}\left(1+\left(\overline{R}_j/(\kappa\underline{R}_j)\right)^{1/\alpha_j}\right)\right]\left[\sum_{j=1}^{d}\max\left\{1,\log\left(\frac{M_j-m_j}{\eta_j}\right)\right\}\right].$$

6. Simulations for multidimensional models

In this section, we complete the simulation study of Section 4 by dealing with multidimensional models.

6.1. Models. We propose to work with the following models.

Example 1. $\mathscr{F}=\left\{f_{(m,\sigma)},\ (m,\sigma)\in[-5,5]\times[1/5,5]\right\}$ *where*

$$f_{(m,\sigma)}(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \quad \textit{for all } x\in\mathbb{R}.$$

Example 2. $\mathscr{F}=\left\{f_{(m,\sigma)},\ (m,\sigma)\in[-5,5]\times[1/5,5]\right\}$ *where*

$$f_{(m,\sigma)}(x)=\frac{\sigma}{\pi\left((x-m)^2+\sigma^2\right)} \quad \textit{for all } x\in\mathbb{R}.$$

Example 3. $\mathscr{F}=\left\{f_{(a,b)},\ (a,b)\in[0.6,10]\times[0.1,20]\right\}$ *where*

$$f_{(a,b)}(x)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}\mathbb{1}_{[0,+\infty)}(x) \quad \textit{for all } x\in\mathbb{R}$$

where Γ is the Gamma function.

Example 4. $\mathscr{F}=\left\{f_{(a,b)},\ (a,b)\in[0.7,20]\times[0.7,20]\right\}$ *where*

$$f_{(a,b)}(x)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x) \quad \textit{for all } x\in\mathbb{R}.$$

and where $B(a,b)$ is the Beta function.

Example 5. $\mathscr{F}=\left\{f_{(m,\lambda)},\ (m,\lambda)\in[-1,1]\times[1/5,5]\right\}$ *where*

$$f_{(m,\lambda)}(x)=\lambda e^{-\lambda(x-m)}\mathbb{1}_{[m,+\infty)}(x) \quad \textit{for all } x\in\mathbb{R}.$$

Example 6. $\mathscr{F} = \{f_{(m,r)},\ (m,r) \in [-0.5, 0.5] \times [0.1, 2]\}$ *where*

$$f_{(m,r)}(x) = r^{-1}\mathbb{1}_{[m,m+r]}(x) \quad \textit{for all } x \in \mathbb{R}.$$

We shall use our procedure with $t_j = 0$ for all $j \in \{1, \dots, d\}$ (that is $\Theta_{\text{dis}} = \Theta$, $\mathscr{F}_{\text{dis}} = \mathscr{F}$ and $T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \overline{T}(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})$) and with $\kappa = 0.9\bar{\kappa}$, $\eta_j = (M_j - m_j)10^{-6}$. In order to avoid technicalities, we delay to Section 8 the values of $\underline{R}_{\mathcal{C},j}$, $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that have been chosen in this simulation study.

6.2. Simulations when $s \in \mathscr{F}$. We simulate $N = 10^4$ independent samples $(X_1, \dots, X_n)$ according to a density $s \in \mathscr{F}$ and use our procedure to estimate s on each of the samples. In examples 1,2,5,6 the density is $s = f_{(0,1)}$, in Example 3, $s = f_{(2,3)}$ and in Example 4, $s = f_{(3,4)}$. The results are the following.

		$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	10^{-4}	$6 \cdot 10^{-5}$	$-5 \cdot 10^{-8}$	$3 \cdot 10^{-8}$
	$\widehat{\text{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.012	0.0055	0.0035	0.0026
	$\widehat{\text{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.012	0.0055	0.0035	0.0026
Example 2	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0052	0.0034	0.0026
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0052	0.0034	0.0026
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	10^{-8}	10^{-8}	-10^{-9}	$4 \cdot 10^{-8}$
	$\widehat{\text{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0052	0.0035	0.0026
	$\widehat{\text{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0052	0.0035	0.0026
Example 3	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-5}$	10^{-7}	10^{-7}
	$\widehat{\text{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0053	0.0035	0.0026
	$\widehat{\text{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0053	0.0035	0.0026
Example 4	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0052	0.0034	0.0025
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	$2 \cdot 10^{-4}$	10^{-5}	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$
	$\widehat{\text{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.011	0.0053	0.0035	0.0026
	$\widehat{\text{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.011	0.0053	0.0035	0.0026
Example 5	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.025	0.012	0.0082	0.0063
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.025	0.012	0.0083	0.0063
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.020	-0.0020	-0.0073	0.0012
	$\widehat{\text{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.025	0.012	0.0079	0.0061
	$\widehat{\text{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.021	0.011	0.0070	0.0053
Example 6	$\widehat{R}_{10^4}(\hat{\boldsymbol{\theta}})$	0.040	0.019	0.013	0.0098
	$\widehat{R}_{10^4}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.039	0.020	0.013	0.010
	$\widehat{\mathcal{R}}_{10^4,\text{rel}}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$	0.010	-0.015	-0.018	-0.016

	$\widehat{\mathrm{std}}_{10^4}(\hat{\boldsymbol{\theta}})$	0.033	0.016	0.011	0.0080
	$\widehat{\mathrm{std}}_{10^4}(\tilde{\boldsymbol{\theta}}_{\mathrm{mle}})$	0.027	0.014	0.0093	0.0069

The risk of our estimator is very close to the one of the m.l.e. In the first four examples they are even almost indistinguishable. As in dimension 1, this can be explained by the fact that the first four models are regular enough to ensure that our estimator is very close to the maximum likelihood one.

To see this, let for $c \in \{0.99, 0.999, 1\}$, q_c be the c-quantile of the random variable

$$\max\left\{|\hat{\theta}_1 - \tilde{\theta}_{\mathrm{mle},1}|, |\hat{\theta}_2 - \tilde{\theta}_{\mathrm{mle},2}|\right\}$$

and $\hat{q}_c$ be the empirical version based on 10^4 samples. These empirical quantiles are very small as shown in the array below.

		$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	$\hat{q}_{0.99}$	$9 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$9 \cdot 10^{-7}$
	$\hat{q}_{0.999}$	0.023	10^{-6}	$9 \cdot 10^{-7}$	10^{-6}
	$\hat{q}_1$	0.22	0.072	10^{-6}	10^{-6}
Example 2	$\hat{q}_{0.99}$	$4 \cdot 10^{-7}$	$4 \cdot 10^{-7}$	$4 \cdot 10^{-7}$	$4 \cdot 10^{-7}$
	$\hat{q}_{0.999}$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$
	$\hat{q}_1$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-7}$
Example 3	$\hat{q}_{0.99}$	$7 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$7 \cdot 10^{-7}$
	$\hat{q}_{0.999}$	$9 \cdot 10^{-7}$	$8 \cdot 10^{-7}$	$8 \cdot 10^{-7}$	$8 \cdot 10^{-7}$
	$\hat{q}_1$	1.5	0.29	10^{-6}	$9 \cdot 10^{-7}$
Example 4	$\hat{q}_{0.99}$	10^{-6}	10^{-6}	10^{-6}	10^{-6}
	$\hat{q}_{0.999}$	$2 \cdot 10^{-6}$	10^{-6}	10^{-6}	10^{-6}
	$\hat{q}_1$	1.6	0.27	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$

6.3. Simulations when $s \notin \mathscr{F}$. Contrary to the maximum likelihood estimator, our estimator possesses robustness properties. The goal of this section is to illustrate them.

Suppose that we observe $n = 100$ i.i.d random variables $X_1, \ldots, X_n$ from which we wish to estimate their distribution by using a Gaussian model

$$\mathscr{F} = \left\{f_{(m,\sigma)},\ (m,\sigma) \in [-10, 10] \times [0.5, 10]\right\}$$

where $f_{(m,\sigma)}$ is the density of a Gaussian random variable with mean m and variance σ^2. The preceding section shows that when the unknown underlying density s belongs to $\mathscr{F}$, our estimator is as good as the m.l.e. We now consider $p \in [0, 1]$ and define $s = s_p$ where

$$s_p(x) = (1 - p)f_{(-5,1)} + pf_{(5,1)} \quad \text{for all } x \in \mathbb{R}.$$

This density belongs to the model only if $p = 0$ or $p = 1$ and we are interested in comparing our estimator to the m.l.e when $p \neq 0$ and $p \neq 1$.

We then proceed as in Section 4.5. For a lot of values of $p \in [0, 1]$, we simulate $N = 1000$ samples of 100 random variables with density s_p and measure the quality of the estimator $\tilde{\boldsymbol{\theta}}$ by

$$\widehat{R}_{p,N}(\tilde{\boldsymbol{\theta}}) = \frac{1}{1000} \sum_{i=1}^{1000} h^2(s_p, f_{\tilde{\boldsymbol{\theta}}^{(p,i)}})$$

where $\tilde{\boldsymbol{\theta}}^{(p,i)}$ is the value of $\tilde{\boldsymbol{\theta}}$ corresponding to the i^{th} sample whose density is s_p. We compute this function for $\tilde{\boldsymbol{\theta}} \in \{\tilde{\boldsymbol{\theta}}_{\text{mle}}, \hat{\boldsymbol{\theta}}\}$ and obtain the graph below.

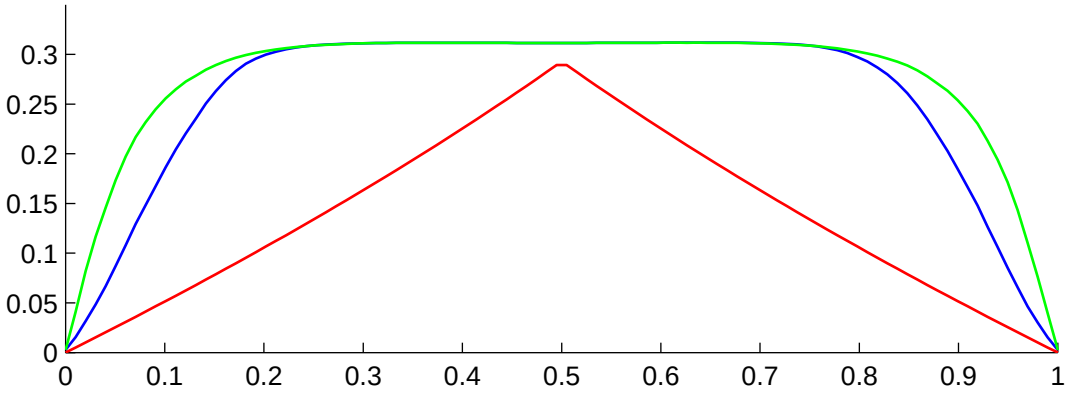


Figure 4.6: Red: $p \mapsto H^2(s_p, \mathscr{F})$. Blue: $p \mapsto \widehat{R}_{p,1000}(\hat{\boldsymbol{\theta}})$. Green: $p \mapsto \widehat{R}_{p,1000}(\tilde{\boldsymbol{\theta}}_{\text{mle}})$.

This figure shows that the risk of our estimator is smaller than the one of the m.l.e when p is close to 0 or 1 (says $p \leq 0.2$ or $p \geq 0.8$) and is similar otherwise. For the Gaussian model, our estimator may thus be interpreted as a robust version of the m.l.e.

7. Proofs

7.1. A preliminary result. In this section, we show a result that will allow us to prove Theorems 3 and 5. Given $\Theta' \subset \Theta$, we recall that $\text{diam}\Theta'$ stands for

$$\text{diam}\Theta' = \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta'} h^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}).$$

Theorem 7. *Suppose that Assumption 1 holds. Let $\kappa \in (0, \bar{\kappa})$, $N \in \mathbb{N}^\star$ and let $\Theta_1 \dots \Theta_N$ be N non-empty subsets of Θ such that $\Theta_1 = \Theta$. For all $j \in \{1, \dots, d\}$, let t_j be an arbitrary number of $(0, d^{1/\alpha_j}]$ and*

$$\epsilon_j = t_j (\overline{R}_j n)^{-1/\alpha_j}.$$

Assume that for all $i \in \{1, \dots, N-1\}$, there exists $L_i \geq 1$ such that for all $\ell \in \{1, \dots, L_i\}$ there

exist two elements $\boldsymbol{\theta}^{(i,\ell)} \neq \boldsymbol{\theta}'^{(i,\ell)}$ *of* Θ_i *such that*

$$\Theta_i \setminus \bigcup_{\ell=1}^{L_i} B^{(i,\ell)} \subset \Theta_{i+1} \subset \Theta_i$$

where $B^{(i,\ell)}$ *is the set defined by*

$$B^{(i,\ell)} = \begin{cases} A^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) > 0 \\ A'^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) < 0 \\ A^{(i,\ell)} \bigcup A'^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) = 0 \end{cases}$$

where $A^{(i,\ell)}$ *and* $A'^{(i,\ell)}$ *are the Hellinger balls defined by*

$$\begin{aligned} A^{(i,\ell)} &= \left\{\boldsymbol{\theta}'' \in \Theta_i,\, h^2(f_{\boldsymbol{\theta}''}, f_{\boldsymbol{\theta}^{(i,\ell)}}) \leq \kappa h^2(f_{\boldsymbol{\theta}^{(i,\ell)}}, f_{\boldsymbol{\theta}'^{(i,\ell)}})\right\} \\ A'^{(i,\ell)} &= \left\{\boldsymbol{\theta}'' \in \Theta_i,\, h^2(f_{\boldsymbol{\theta}''}, f_{\boldsymbol{\theta}'^{(i,\ell)}}) \leq \kappa h^2(f_{\boldsymbol{\theta}^{(i,\ell)}}, f_{\boldsymbol{\theta}'^{(i,\ell)}})\right\} \end{aligned}$$

and where T *is the functional defined by (14). Suppose moreover that there exists* $\kappa_0 > 0$ *such that*

$$\kappa_0 \, diam(\Theta_i) \leq \inf_{1 \leq \ell \leq L_i} h^2(f_{\boldsymbol{\theta}^{(i,\ell)}}, f_{\boldsymbol{\theta}'^{(i,\ell)}}) \quad \text{for all } i \in \{1, \ldots, N-1\}.$$

Then, for all $\xi > 0$,

$$\mathbb{P}\left[C \inf_{\boldsymbol{\theta} \in \Theta_N} h^2(s, f_{\boldsymbol{\theta}}) \geq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi\right] \leq e^{-n\xi}$$

where $C > 0$ *depends only on* κ, κ_0, *where*

$$D_{\mathscr{F}} = \max\left\{d, \sum_{j=1}^{d} \log\left(1 + t_j^{-1}\left((d/\bar{\boldsymbol{\alpha}})(c\overline{R}_j/\underline{R}_j)\right)^{1/\alpha_j}\right)\right\}$$

and where c *depends only on* κ.

This result says that if $(\Theta_i)_{1 \leq i \leq N}$ is a finite sequence of subsets of Θ satisfying the assumptions of the theorem, then there exists an estimator $\hat{\boldsymbol{\theta}}$ with values in Θ_N whose risk can be upper bounded by

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\boldsymbol{\theta}}}) \geq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi\right] \leq e^{-n\xi}.$$

We shall show that algorithms 1 and 3 correspond to suitable choices of sets Θ_i.

7.2. Proof of Theorem 7. Let $\boldsymbol{\theta}_0 \in \Theta$ be such that

$$h^2(s, f_{\boldsymbol{\theta}_0}) \leq h^2(s, \mathscr{F}) + 1/n.$$

Define C_κ such that $(1 + \sqrt{C_\kappa})^2 = \kappa^{-1}$ and $\varepsilon \in (1/\sqrt{2}, 1)$ such that

$$\frac{\left(1 + \min\left(\frac{1-\varepsilon}{2}, \varepsilon - \frac{1}{\sqrt{2}}\right)\right)^4 (1+\varepsilon) + \min\left(\frac{1-\varepsilon}{2}, \varepsilon - \frac{1}{\sqrt{2}}\right)}{1 - \varepsilon - \min\left(\frac{1-\varepsilon}{2}, \varepsilon - \frac{1}{\sqrt{2}}\right)} = C_\kappa.$$

We then set

$$\begin{aligned}
\beta &= \min\left\{(1-\varepsilon)/2, \varepsilon - 1/\sqrt{2}\right\} \\
\gamma &= (1+\beta)(1+\beta^{-1})\left[1-\varepsilon+(1+\beta)(1+\varepsilon)\right] \\
c &= 24\big(2+\sqrt{2}/6\left(\varepsilon-1/\sqrt{2}\right)\big)/(\varepsilon-1/\sqrt{2})^2\cdot 10^3 \\
\delta &= (1+\beta^{-1})\left[1-\varepsilon+(1+\beta)^3(1+\varepsilon)\right] + c(1+\beta)^2.
\end{aligned}$$

The proof of the theorem is based on the lemma below whose proof is delayed to Section 7.2.1.

Lemma 1. *For all $\xi > 0$, there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which the following assertion holds: if there exists $p \in \{1, \ldots, N-1\}$ such that $\boldsymbol{\theta}_0 \in \Theta_p$ and such that*

$$\gamma h^2(s, f_{\boldsymbol{\theta}_0}) + \delta\left(\frac{D_{\mathscr{F}}}{n} + \xi\right) < \beta \inf_{\ell\in\{1,\ldots,L_p\}} \left(h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}})\right) \tag{21}$$

then $\boldsymbol{\theta}_0 \in \Theta_{p+1}$.

The result of Theorem 7 is straightforward if $\boldsymbol{\theta}_0 \in \Theta_N$, and we shall thus assume that $\boldsymbol{\theta}_0 \notin \Theta_N$. Set

$$p = \max\left\{i \in \{1, \ldots, N-1\},\, \boldsymbol{\theta}_0 \in \Theta_i\right\}.$$

Let $\boldsymbol{\theta}'_0$ be any element of Θ_N. Then, $\boldsymbol{\theta}'_0$ belongs to Θ_p and

$$\begin{aligned}
h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'_0}) &\leq \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta_p} h^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}) \\
&\leq \kappa_0^{-1} \inf_{\ell\in\{1,\ldots,L_p\}} h^2(f_{\boldsymbol{\theta}^{(p,\ell)}}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) \\
&\leq 2\kappa_0^{-1} \inf_{\ell\in\{1,\ldots,L_p\}} \left(h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}})\right).
\end{aligned}$$

By definition of p, $\boldsymbol{\theta}_0 \in \Theta_p \setminus \Theta_{p+1}$. We then derive from the above lemma that on Ω_ξ,

$$\beta \inf_{\ell\in\{1,\ldots,L_p\}} \left(h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}})\right) \leq \gamma h^2(s, f_{\boldsymbol{\theta}_0}) + \delta\frac{D_{\mathscr{F}} + n\xi}{n}.$$

Hence,

$$h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'_0}) \leq \frac{2}{\beta\kappa_0}\left(\gamma h^2(s, f_{\boldsymbol{\theta}_0}) + \delta\frac{D_{\mathscr{F}} + n\xi}{n}\right)$$

and thus

$$\begin{aligned}
h^2(s, f_{\boldsymbol{\theta}'_0}) &\leq 2h^2\left(s, f_{\boldsymbol{\theta}_0}\right) + 2h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'_0}) \\
&\leq \left(2 + \frac{4\gamma}{\beta\kappa_0}\right) h^2(s, f_{\boldsymbol{\theta}_0}) + \frac{4\delta}{n\beta\kappa_0}\left(D_{\mathscr{F}} + n\xi\right).
\end{aligned}$$

Since $h^2(s, f_{\boldsymbol{\theta}_0}) \leq h^2(s, \mathscr{F}) + 1/n$, there exists $C > 0$ such that

$$Ch^2(s, f_{\boldsymbol{\theta}'_0}) \leq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi \quad \text{on } \Omega_\xi.$$

This concludes the proof. □

7.2.1. Proof of Lemma 1. We use the claim below whose proof is postponed to Section 7.2.2

Claim 1. *For all $\xi > 0$, there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which, for all $f, f' \in \mathscr{F}_{dis}$,*

$$(1-\varepsilon)\, h^2(s,f') + \frac{\overline{T}(f,f')}{\sqrt{2}} \leq (1+\varepsilon)\, h^2(s,f) + c\frac{(D_{\mathscr{F}} + n\xi)}{n}$$

(see Section 5.2 for the definition of $\mathscr{F}_{dis}$).

Let $p \in \{1,\dots,N-1\}$ such that $\boldsymbol{\theta}_0 \in \Theta_p$ and (21) holds. Let $\ell \in \{1,\dots,L_p\}$. The aim is to show that $\boldsymbol{\theta}_0 \notin B^{(p,\ell)}$. Without lost of generality, we assume that $T(\boldsymbol{\theta}^{(p,\ell)}, \boldsymbol{\theta}'^{(p,\ell)}) = \overline{T}(f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})})$ is non-negative, and prove that $\boldsymbol{\theta}_0 \notin A^{(p,\ell)}$.

On the event Ω_ξ, we deduce from the claim

$$(1-\varepsilon)\, h^2(s, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})}) \leq (1+\varepsilon)\, h^2(s, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}) + c\frac{(D_{\mathscr{F}} + n\xi)}{n}.$$

Consequently, by using the triangular inequality and the above inequality

$$\begin{aligned}
(1-\varepsilon)\, h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})}) &\leq \left(1+\beta^{-1}\right)(1-\varepsilon)\, h^2(s, f_{\boldsymbol{\theta}_0}) \\
&\quad +(1+\beta)(1-\varepsilon)\, h^2(s, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})}) \\
&\leq \left(1+\beta^{-1}\right)(1-\varepsilon)\, h^2(s, f_{\boldsymbol{\theta}_0}) \\
&\quad +(1+\beta)\left[(1+\varepsilon)\, h^2(s, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}) + c\frac{(D_{\mathscr{F}} + n\xi)}{n}\right].
\end{aligned}$$

Since $h^2(s, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}) \leq (1+\beta^{-1})h^2(s, f_{\boldsymbol{\theta}_0}) + (1+\beta)h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})})$,

$$\begin{aligned}
(1-\varepsilon)\, h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})}) &\leq (1+\beta^{-1})\left[1-\varepsilon+(1+\beta)(1+\varepsilon)\right] h^2(s, f_{\boldsymbol{\theta}_0}) \\
&\quad +(1+\beta)^2(1+\varepsilon)h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}) \\
&\quad +\frac{c(1+\beta)\,(D_{\mathscr{F}} + n\xi)}{n}.
\end{aligned} \tag{22}$$

Remark now that for all $\boldsymbol{\theta} \in \Theta$,

$$h^2(f_{\boldsymbol{\theta}}, f_{\pi(\boldsymbol{\theta})}) \leq \sup_{1\leq j\leq d} \overline{R}_j \epsilon_j^{\alpha_j} \leq d/n.$$

By using the triangular inequality,

$$\begin{aligned}
h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}^{(p,\ell)})}) &\leq (1+\beta)h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + d(1+\beta^{-1})/n \\
h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) &\leq (1+\beta)h^2(f_{\boldsymbol{\theta}_0}, f_{\pi(\boldsymbol{\theta}'^{(p,\ell)})}) + d(1+\beta^{-1})/n.
\end{aligned}$$

We deduce from these two inequalities and from (22) that

$$\begin{aligned}
(1-\varepsilon)\, h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) \leq & \gamma h^2(s, f_{\boldsymbol{\theta}_0}) + (1+\beta)^4(1+\varepsilon)h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) \\
&+\frac{d(1+\beta^{-1})\left[1-\varepsilon+(1+\beta)^3(1+\varepsilon)\right] + c(1+\beta)^2\,(D_{\mathscr{F}} + n\xi)}{n}.
\end{aligned}$$

Since $D_{\mathscr{F}} \geq d$ and $\delta \geq 1$

$$\begin{aligned}(1-\varepsilon)\, h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) &\leq \gamma h^2(s, f_{\boldsymbol{\theta}_0}) + \frac{\delta\,(D_{\mathscr{F}} + n\xi)}{n} \\ &\quad +(1+\beta)^4(1+\varepsilon)h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}).\end{aligned}$$

By using (21),

$$\begin{aligned}(1-\varepsilon)\, h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) &< \beta\left(h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}})\right) \\ &\quad +(1+\beta)^4(1+\varepsilon)h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}})\end{aligned}$$

and thus

$$h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) < C_\kappa h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}).$$

Finally,

$$\begin{aligned}h^2(f_{\boldsymbol{\theta}^{(p,\ell)}}, f_{\boldsymbol{\theta}'^{(p,\ell)}}) &\leq \left(h(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) + h(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}'^{(p,\ell)}})\right)^2 \\ &< \left(1+\sqrt{C_\kappa}\right)^2 h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}}) \\ &< \kappa^{-1} h^2(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}^{(p,\ell)}})\end{aligned}$$

which leads to $\boldsymbol{\theta}_0 \notin A^{(p,\ell)}$ as wished.

7.2.2. Proof of Claim 1. This claim ensues from the work of Baraud (2011). More precisely, we derive from Proposition 2 of Baraud (2011) that for all $f, f' \in \mathscr{F}_{\text{dis}}$,

$$\left(1-\frac{1}{\sqrt{2}}\right) h^2(s,f') + \frac{\overline{T}(f,f')}{\sqrt{2}} \leq \left(1+\frac{1}{\sqrt{2}}\right) h^2(s,f) + \frac{\overline{T}(f,f') - \mathbb{E}\left[\overline{T}(f,f')\right]}{\sqrt{2}}.$$

Let $z = \varepsilon - 1/\sqrt{2} \in (0, 1-1/\sqrt{2})$. We define Ω_ξ by

$$\Omega_\xi = \bigcap_{f,f' \in \mathscr{F}_{\text{dis}}} \left[\frac{\overline{T}(f,f') - \mathbb{E}\left[\overline{T}(f,f')\right]}{z\left(h^2(s,f) + h^2(s,f')\right) + c(D_{\mathscr{F}} + n\xi)/n} \leq \sqrt{2}\right].$$

On this event, we have

$$(1-\varepsilon)\, h^2(s,f') + \frac{\overline{T}(f,f')}{\sqrt{2}} \leq (1+\varepsilon)\, h^2(s,f) + c\frac{D_{\mathscr{F}} + n\xi}{n}$$

and it remains to prove that $\mathbb{P}\left(\Omega_\xi^c\right) \leq e^{-n\xi}$.

The following claim shows that Assumption 3 of Baraud (2011) is fulfilled.

Claim 2. *Let*

$$\begin{aligned}\tau &= 4\frac{2 + \frac{n\sqrt{2}}{6}z}{\frac{n^2}{6}z^2} \\ \eta_{\mathscr{F}}^2 &= \max\left\{3de^4, \sum_{j=1}^{d} \log\left(1 + 2t_j^{-1}\left((d/\bar{\boldsymbol{\alpha}})(c\overline{R}_j/\underline{R}_j)\right)^{1/\alpha_j}\right)\right\}.\end{aligned}$$

Then, for all $r \geq 2\eta_{\mathscr{F}}$,

$$\left|\mathscr{F}_{dis} \cap \mathcal{B}_h\left(s, r\sqrt{\tau}\right)\right| \leq \exp(r^2/2) \tag{23}$$

where $\mathcal{B}_h(s, r\sqrt{\tau})$ *is the Hellinger ball centered at* s *with radius* $r\sqrt{\tau}$ *defined by*

$$\mathcal{B}_h(s, r\sqrt{\tau}) = \left\{f \in \mathbb{L}_+^1(\mathbb{X}, \mu),\ h^2(s, f) \leq r^2\tau\right\}.$$

We then derive from Lemma 1 of Baraud (2011) that for all $\xi > 0$ and $y^2 \geq \tau\left(4\eta_{\mathscr{F}}^2 + n\xi\right)$,

$$\mathbb{P}\left[\sup_{f,f' \in \mathscr{F}_{\text{dis}}} \frac{\left(\overline{T}(f, f') - \mathbb{E}\left[\overline{T}(f, f')\right]\right)/\sqrt{2}}{(h^2(s, f) + h^2(s, f')) \vee y^2} \geq z\right] \leq e^{-n\xi}.$$

Notice now that $4\eta_{\mathscr{F}}^2 \leq 10^3 D_{\mathscr{F}}$ and $10^3\tau \leq c/n$. This means that we can choose

$$y^2 = c\left(D_{\mathscr{F}} + n\xi\right)/n,$$

which concludes the proof of Claim 1.

Proof of Claim 2. If $\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(s, r\sqrt{\tau}) = \emptyset$, (23) holds. In the contrary case, there exists $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,d}) \in \Theta_{\text{dis}}$ such that $h^2(s, f_{\boldsymbol{\theta}_0}) \leq r^2\tau$ and thus

$$|\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(s, r\sqrt{\tau})| \leq |\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})|.$$

Now,

$$\begin{aligned} |\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})| &= \left|\left\{f_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta_{\text{dis}},\ h^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}_0}) \leq 4r^2\tau\right\}\right| \\ &\leq \left|\left\{\boldsymbol{\theta} \in \Theta_{\text{dis}},\ \forall j \in \{1, \ldots, d\},\ \underline{R}_j|\theta_j - \theta_{0,j}|^{\alpha_j} \leq 4r^2\tau\right\}\right|. \end{aligned}$$

Let $k_{0,j} \in \mathbb{N}$ be such that $\theta_{0,j} = m_j + k_{0,j}\epsilon_j$. Then,

$$\begin{aligned} |\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})| &\leq \prod_{j=1}^{d}\left|\left\{k_j \in \mathbb{N},\ |k_j - k_{0,j}| \leq \left(4r^2\tau/\underline{R}_j\right)^{1/\alpha_j}\epsilon_j^{-1}\right\}\right| \\ &\leq \prod_{j=1}^{d}\left(1 + 2\epsilon_j^{-1}\left(4r^2\tau/\underline{R}_j\right)^{1/\alpha_j}\right). \end{aligned}$$

By using $4\tau \leq c/n$ and $\epsilon_j = t_j(\overline{R}_j n)^{-1/\alpha_j}$,

$$|\mathscr{F}_{\text{dis}} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})| \leq \prod_{j=1}^{d}\left(1 + 2t_j^{-1}\left(r^2 c\overline{R}_j/\underline{R}_j\right)^{1/\alpha_j}\right).$$

If $\bar{\boldsymbol{\alpha}} \leq e^{-4}$, one can check that $\eta_{\mathscr{F}}^2 \geq 4d/\bar{\boldsymbol{\alpha}}$ (since $c \geq 1$ and $t_j^{-1} \geq d^{-1/\alpha_j}$). If now $\bar{\boldsymbol{\alpha}} \geq e^{-4}$, then $\eta_{\mathscr{F}}^2 \geq 3de^4 \geq 3d/\bar{\boldsymbol{\alpha}}$. In particular, we always have $r^2 \geq 10\,(d/\bar{\boldsymbol{\alpha}})$.

We derive from the weaker inequality $r^2 \geq d/\bar{\boldsymbol{\alpha}}$ that

$$\begin{aligned} |\mathscr{F}_{\rm dis} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})| &\leq \left(\frac{r^2}{d/\bar{\boldsymbol{\alpha}}}\right)^{d/\bar{\boldsymbol{\alpha}}} \prod_{j=1}^{d} \left(1 + 2t_j^{-1}\left((d/\bar{\boldsymbol{\alpha}})(c\overline{R}_j/\underline{R}_j)\right)^{1/\alpha_j}\right) \\ &\leq \exp\left(\frac{\log\left(r^2/(d/\bar{\boldsymbol{\alpha}})\right)}{r^2/(d/\bar{\boldsymbol{\alpha}})} r^2\right) \exp\left(\eta_{\mathscr{F}}^2\right). \end{aligned}$$

We then deduce from the inequalities $r^2/(d/\bar{\boldsymbol{\alpha}}) \geq 10$ and $\eta_{\mathscr{F}}^2 \leq r^2/4$ that

$$|\mathscr{F}_{\rm dis} \cap \mathcal{B}_h(f_{\boldsymbol{\theta}_0}, 2r\sqrt{\tau})| \leq \exp\left(r^2/4\right) \exp\left(r^2/4\right) \leq \exp(r^2/2)$$

as wished. □

7.3. Proof of Theorem 3. This theorem ensues from the following result.

Theorem 8. *Suppose that the assumptions of Theorem 3 holds. For all $\xi > 0$, the estimator $\hat{\theta}$ built in Algorithm 1 satisfies*

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \geq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi\right] \leq e^{-n\xi}$$

where $C > 0$ depends only on $\kappa, \overline{R}/\underline{R}$, where

$$D_{\mathscr{F}} = \max\left\{1, \log\left(1 + t^{-1}\big(c\overline{R}/(\alpha\underline{R})\big)^{1/\alpha}\right)\right\}$$

and where c depends on κ only. Besides, if

$$h^2(f_{\theta_2}, f_{\theta_2'}) \leq h^2(f_{\theta_1}, f_{\theta_1'}) \quad \textit{for all } m \leq \theta_1 \leq \theta_2 < \theta_2' \leq \theta_1' \leq M$$

then C depends only on κ.

Proof. The theorem follows from Theorem 7 page 134 where $\Theta_i = [\theta^{(i)}, \theta'^{(i)}]$ and $L_i = 1$. Note that

$$\mathrm{diam}\Theta_i \leq \overline{R}\big(\theta'^{(i)} - \theta^{(i)}\big)^{\alpha} \leq (\overline{R}/\underline{R}) h^2(f_{\theta^{(i)}}, f_{\theta'^{(i)}})$$

which implies that the assumptions of Theorem 7 are fulfilled with $\kappa_0 = \underline{R}/\overline{R}$. Consequently,

$$\mathbb{P}\left[C \inf_{\theta \in \Theta_N} h^2(s, f_\theta) \geq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi\right] \leq e^{-n\xi}$$

where $\Theta_N = [\theta^{(N)}, \theta'^{(N)}]$ is such that $\theta'^{(N)} - \theta^{(N)} \leq \eta$. Now, for all $\theta \in \Theta_N$,

$$\begin{aligned} h^2(s, f_{\hat{\theta}}) &\leq 2h^2(s, f_\theta) + 2h^2(f_\theta, f_{\hat{\theta}}) \\ &\leq 2h^2(s, f_\theta) + 2\overline{R}\eta^{\alpha} \end{aligned}$$

hence,

$$h^2(s, f_{\hat{\theta}}) \leq 2 \inf_{\theta \in \Theta_N} h^2(s, f_\theta) + 2/n$$

which establishes the first part of the theorem. The second part derives from the fact that under the additional assumption, $\mathrm{diam}\Theta_i \leq h^2(f_{\theta^{(i)}}, f_{\theta'^{(i)}})$, which means that the assumptions of Theorem 7 are fulfilled with $\kappa_0 = 1$. □

7.4. Proof of Proposition 4. For all $i \in \{1, \dots, N-1\}$,

$$\begin{aligned} \theta^{(i+1)} &\in \left\{\theta^{(i)}, \theta^{(i)} + \min\left(\bar{r}(\theta^{(i)}, \theta'^{(i)}), (\theta'^{(i)} - \theta^{(i)})/2\right)\right\} \\ \theta'^{(i+1)} &\in \left\{\theta'^{(i)}, \theta'^{(i)} - \min\left(\underline{r}(\theta^{(i)}, \theta'^{(i)}), (\theta'^{(i)} - \theta^{(i)})/2\right)\right\}. \end{aligned}$$

Since $\bar{r}(\theta^{(i)}, \theta'^{(i)})$ and $\underline{r}(\theta^{(i)}, \theta'^{(i)})$ are larger than

$$(\kappa \underline{R}/\overline{R})^{1/\alpha}(\theta'^{(i)} - \theta^{(i)}),$$

we have

$$\theta'^{(i+1)} - \theta^{(i+1)} \leq \max\left\{1 - \left(\kappa \underline{R}/\overline{R}\right)^{1/\alpha}, 1/2\right\}(\theta'^{(i)} - \theta^{(i)}).$$

By induction, we derive that for all $i \in \{1, \dots, N-1\}$,

$$\theta'^{(i+1)} - \theta^{(i+1)} \leq \left(\max\left\{1 - \left(\kappa \underline{R}/\overline{R}\right)^{1/\alpha}, 1/2\right\}\right)^{i}(M - m).$$

The procedure requires thus the computation of at most N tests where N is the smallest integer such that

$$\left(\max\left\{1 - \left(\kappa \underline{R}/\overline{R}\right)^{1/\alpha}, 1/2\right\}\right)^{N}(M - m) \leq \eta$$

that is

$$N \geq \frac{\log\left((M-m)/\eta\right)}{-\log\left[\max\left\{1 - \left(\kappa \underline{R}/\overline{R}\right)^{1/\alpha}, 1/2\right\}\right]}.$$

We conclude by using the inequality $-1/\log(1-x) \leq 1/x$ for all $x \in (0,1)$. □

7.5. Proofs of Proposition 6 and Theorem 5.

7.5.1. Rewriting of Algorithm 3. We rewrite the algorithm to introduce some notations that will be essential to prove Proposition 6 and Theorem 5.

Algorithm 4 Construction of Θ_{i+1} from Θ_i.

Require: $\Theta_i = \prod_{j=1}^{d}[a_j^{(i)}, b_j^{(i)}]$

1: Choose $k^{(i)} \in \{1, \dots, d\}$ such that

$$\underline{R}_{\Theta_i, k^{(i)}}\left(b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)}\right)^{\alpha_{k^{(i)}}} = \max_{1 \leq j \leq d} \underline{R}_{\Theta_i, j}\left(b_j^{(i)} - a_j^{(i)}\right)^{\alpha_j}.$$

2: $\boldsymbol{\theta}^{(i,1)} = (a_1^{(i)}, \dots, a_d^{(i)})$, $\boldsymbol{\theta}'^{(i,1)} = \boldsymbol{\theta}^{(i,1)}$ and $\theta_{k^{(i)}}'^{(i,1)} = b_{k^{(i)}}^{(i)}$.

3: ${\varepsilon_j}^{(i,0)} = \bar{r}_{\Theta_i, j}(\boldsymbol{\theta}^{(i,1)}, \boldsymbol{\theta}'^{(i,1)})$ and ${\varepsilon_j}'^{(i,0)} = \bar{r}_{\Theta_i, j}(\boldsymbol{\theta}'^{(i,1)}, \boldsymbol{\theta}^{(i,1)})$ for all $j \neq k^{(i)}$

4: $\varepsilon_{k^{(i)}}^{(i,0)} = (b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)})/2$ and $\varepsilon_{k^{(i)}}'^{(i,0)} = (b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)})/2$

for all $\ell \geq 1$ **do**
  $\boldsymbol{\theta}^{(i,\ell+1)} = \boldsymbol{\theta}^{(i,\ell)}$ and $\boldsymbol{\theta}'^{(i,\ell+1)} = \boldsymbol{\theta}'^{(i,\ell)}$
  if $T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \geq 0$ **then**
    $\varepsilon^{(i,\ell)}_{\psi_{k^{(i)}}(1)} = \bar{r}_{\Theta_i, \psi_{k^{(i)}}(1)}(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)})$
    $\varepsilon^{(i,\ell)}_{\psi_{k^{(i)}}(j)} = \min(\varepsilon^{(i,\ell-1)}_{\psi_{k^{(i)}}(j)}, \bar{r}_{\Theta_i, \psi_{k^{(i)}}(j)}(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}))$, for all $j \in \{2, \ldots, d-1\}$
    $\varepsilon^{(i,\ell)}_{k^{(i)}} = \min(\varepsilon^{(i,\ell-1)}_{k^{(i)}}, \bar{r}_{\Theta_i, k^{(i)}}(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}))$
    $\mathfrak{J}^{(i,\ell)} = \left\{1 \leq j \leq d-1,\ \theta^{(i,\ell)}_{\psi_{k^{(i)}}(j)} + \varepsilon^{(i,\ell)}_{\psi_{k^{(i)}}(j)} < b^{(i)}_{\psi_{k^{(i)}}(j)}\right\}$
    if $\mathfrak{J}^{(i,\ell)} \neq \emptyset$ **then**
      $\mathrm{j}^{(i,\ell)}_{\min} = \min \mathfrak{J}^{(i,\ell)}$
      $\theta^{(i,\ell+1)}_{\psi_{k^{(i)}}(j)} = a^{(i)}_{\psi_{k^{(i)}}(j)}$ for all $j \leq \mathrm{j}^{(i,\ell)}_{\min} - 1$
      $\theta^{(i,\ell+1)}_{\psi_{k^{(i)}}(\mathrm{j}^{(i,\ell)}_{\min})} = \theta^{(i,\ell)}_{\psi_{k^{(i)}}(\mathrm{j}^{(i,\ell)}_{\min})} + \varepsilon^{(i,\ell)}_{\psi_{k^{(i)}}(\mathrm{j}^{(i,\ell)}_{\min})}$
    else
      $\mathrm{j}^{(i,\ell)}_{\min} = d$
    end if
  end if
  if $T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \leq 0$ **then**
    $\varepsilon'^{(i,\ell)}_{\psi_{k^{(i)}}(1)} = \bar{r}_{\Theta_i, \psi_{k^{(i)}}(1)}(\boldsymbol{\theta}'^{(i,\ell)}, \boldsymbol{\theta}^{(i,\ell)})$
    $\varepsilon'^{(i,\ell)}_{\psi_{k^{(i)}}(j)} = \min(\varepsilon'^{(i,\ell-1)}_{\psi_{k^{(i)}}(j)}, \bar{r}_{\Theta_i, \psi_{k^{(i)}}(j)}(\boldsymbol{\theta}'^{(i,\ell)}, \boldsymbol{\theta}^{(i,\ell)}))$, for all $j \in \{2, \ldots, d-1\}$
    $\varepsilon'^{(i,\ell)}_{k^{(i)}} = \min(\varepsilon'^{(i,\ell-1)}_{k^{(i)}}, \underline{r}_{\Theta_i, k^{(i)}}(\boldsymbol{\theta}'^{(i,\ell)}, \boldsymbol{\theta}^{(i,\ell)}))$
    $\mathfrak{J}'^{(i,\ell)} = \left\{1 \leq j \leq d-1,\ \theta'^{(i,\ell)}_{\psi_{k^{(i)}}(j)} + \varepsilon'^{(i,\ell)}_{\psi_{k^{(i)}}(j)} < b^{(i)}_{\psi_{k^{(i)}}(j)}\right\}$
    if $\mathfrak{J}'^{(i,\ell)} \neq \emptyset$ **then**
      $\mathrm{j}'^{(i,\ell)}_{\min} = \mathfrak{J}'^{(i,\ell)}$
      $\theta'^{(i,\ell+1)}_{\psi_{k^{(i)}}(j)} = a^{(i)}_{\psi_{k^{(i)}}(j)}$ for all $j \leq \mathrm{j}'^{(i,\ell)}_{\min} - 1$
      $\theta'^{(i,\ell+1)}_{\psi_{k^{(i)}}(\mathrm{j}'^{(i,\ell)}_{\min})} = \theta^{(i,\ell)}_{\psi_{k^{(i)}}(\mathrm{j}'^{(i,\ell)}_{\min})} + \varepsilon'^{(i,\ell)}_{\psi_{k^{(i)}}(\mathrm{j}'^{(i,\ell)}_{\min})}$
    else
      $\mathrm{j}'^{(i,\ell)}_{\min} = d$
    end if
  end if
  if $\mathrm{j}^{(i,\ell)}_{\min} = d$ or $\mathrm{j}'^{(i,\ell)}_{\min} = d$ **then**
    $L_i = \ell$ and quit the loop
  end if
  $a^{(i+1)}_j = a^{(i)}_j$ and $b^{(i+1)}_j = b^{(i)}_j$ for all $j \neq k^{(i)}$
end for
if $\mathrm{j}^{(i,\ell)}_{\min} = d$ **then**
  $a^{(i+1)}_{k^{(i)}} = a^{(i)}_{k^{(i)}} + \varepsilon^{(i,L_i)}_{k^{(i)}}$
end if
if $\mathrm{j}'^{(i,\ell)}_{\min} = d$ **then**
  $b^{(i+1)}_{k^{(i)}} = b^{(i)}_{k^{(i)}} - \varepsilon'^{(i,L_i)}_{k^{(i)}}$
end if
Return: $\Theta_{i+1} = \prod_{j=1}^{d} [a^{(i+1)}_j, b^{(i+1)}_j]$

Algorithm 4 Rewriting of Algorithm 3.

$\Theta_1 = \prod_{j=1}^d [a_j^{(1)}, b_j^{(1)}] = \prod_{j=1}^d [m_j, M_j]$
for all $i \geq 1$ **do**
  if There exists $j \in \{1, \ldots, d\}$ such that $b_j^{(i)} - a_j^{(i)} > \eta_j$ **then**
    Compute Θ_{i+1}
  else
    Leave the loop and set $N = i$
  end if
end for
Return:

$$\hat{\boldsymbol{\theta}} = \left(\frac{a_1^{(N)} + b_1^{(N)}}{2}, \ldots, \frac{a_d^{(N)} + b_d^{(N)}}{2} \right)$$

7.5.2. Proof of Proposition 6. The algorithm computes $\sum_{i=1}^N L_i$ tests. Define for all $j \in \{1, \ldots, d\}$,

$$I_j = \left\{ i \in \{1, \ldots, N\},\, k^{(i)} = j \right\}.$$

Then, $\cup_{j=1}^d I_j = \{1, \ldots, N\}$. Since

$$\sum_{i=1}^N L_i \leq \sum_{j=1}^d |I_j| \sup_{i \in I_j} L_i, \tag{24}$$

we begin to bound $|I_j|$ from above. For all $i \in \{1, \ldots, N-1\}$,

$$\begin{aligned} b_j^{(i+1)} - a_j^{(i+1)} &\leq b_j^{(i)} - a_j^{(i)} - \min\left(\varepsilon_j^{(i,L_i)}, \varepsilon_j'^{(i,L_i)} \right) \quad \text{if } i \in I_j \\ b_j^{(i+1)} - a_j^{(i+1)} &= b_j^{(i)} - a_j^{(i)} \quad \text{if } i \notin I_j. \end{aligned}$$

For all $i \in I_j$, and all $\ell \in \{1, \ldots, L_i\}$, we derive from (20), from the equality $\theta_j'^{(i,\ell)} = b_j^{(i)}$, $\theta_j^{(i,\ell)} = a_j^{(i)}$ and from the inequalities $\underline{R}_{\Theta_i,j} \geq \underline{R}_j$ and $\overline{R}_{\Theta_i,j} \leq \overline{R}_j$ that

$$\bar{r}_{\Theta_i,j}(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \geq (\kappa \underline{R}_j / \overline{R}_j)^{1/\alpha_j} (b_j^{(i)} - a_j^{(i)}) \tag{25}$$

$$\underline{r}_{\Theta_i,j}(\boldsymbol{\theta}'^{(i,\ell)}, \boldsymbol{\theta}^{(i,\ell)}) \geq (\kappa \underline{R}_j / \overline{R}_j)^{1/\alpha_j} (b_j^{(i)} - a_j^{(i)}). \tag{26}$$

Consequently,

$$\min\left(\varepsilon_j^{(i,L_i)}, \varepsilon_j'^{(i,L_i)} \right) \geq \min\left\{ (b_j^{(i)} - a_j^{(i)})/2, (\kappa \underline{R}_j / \overline{R}_j)^{1/\alpha_j} (b_j^{(i)} - a_j^{(i)}) \right\}.$$

We then have,

$$\begin{aligned} b_j^{(i+1)} - a_j^{(i+1)} &\leq \max\left(1/2, 1 - \left(\kappa \underline{R}_j / \overline{R}_j \right)^{1/\alpha_j} \right) \left(b_j^{(i)} - a_j^{(i)} \right) \quad \text{when } i \in I_j \\ b_j^{(i+1)} - a_j^{(i+1)} &= b_j^{(i)} - a_j^{(i)} \quad \text{when } i \notin I_j. \end{aligned}$$

Let n_j be any integer such that

$$\left(\max\left\{1/2, 1-\left(\kappa\underline{R}_j/\overline{R}_j\right)^{1/\alpha_j}\right\}\right)^{n_J} \leq \eta_j/(M_j-m_j).$$

If $|I_j| > n_j$, then for $i = \max I_j$,

$$\begin{aligned} b_j^{(i)} - a_j^{(i)} &\leq \left(\max\left\{1/2, 1-\left(\kappa\underline{R}_j/\overline{R}_j\right)^{1/\alpha_j}\right\}\right)^{|I_j|-1}(M_j-m_j) \\ &\leq \left(\max\left\{1/2, 1-\left(\kappa\underline{R}_j/\overline{R}_j\right)^{1/\alpha_j}\right\}\right)^{n_J}(M_j-m_j) \end{aligned}$$

and thus $b_j^{(i)} - a_j^{(i)} \leq \eta_j$. This is impossible because $i \in I_j$ implies that $b_j^{(i)} - a_j^{(i)} > \eta_j$. Consequently, $|I_J| \leq n_j$. We then set n_j as the smallest integer larger than

$$\frac{\log\left(\eta_j/(M_j-m_j)\right)}{\log\left(\max\left\{1/2, 1-\left(\kappa\underline{R}_j/\overline{R}_j\right)^{1/\alpha_j}\right\}\right)}.$$

By using the inequality $-1/\log(1-x) \leq 1/x$ for all $x \in (0,1)$, we obtain

$$|I_j| \leq 1 + \max\left(1/\log 2, \left(\kappa\underline{R}_j/\overline{R}_j\right)^{-1/\alpha_j}\right)\log\left(\frac{M_j-m_j}{\eta_j}\right).$$

We now roughly bound from above the right-hand side of this inequality:

$$|I_j| \leq 2\left(1+\left(\overline{R}_j/(\kappa\underline{R}_j)\right)^{1/\alpha_j}\right)\left(1 \vee \log\left(\frac{M_j-m_j}{\eta_j}\right)\right).$$

We recall that our aim is to bound from above $\sum_{i=1}^N L_i$. Thanks to (24), it remains to upper bound $\sup_{i\in I_j} L_i$. This ensues from the following lemma.

Lemma 2. *Let*

$$\mathcal{L} = \left\{1 \leq \ell \leq L_i,\ T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \geq 0\right\} \quad \textit{and} \quad \mathcal{L}' = \left\{1 \leq \ell \leq L_i,\ T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \leq 0\right\}.$$

Then,

$$\begin{aligned} |\mathcal{L}| &\leq \prod_{k\in\{1,\ldots,d\}\setminus\{k^{(i)}\}}\left[1+\left(\overline{R}_k/(\kappa\underline{R}_k)\right)^{1/\alpha_k}\right] \\ |\mathcal{L}'| &\leq \prod_{k\in\{1,\ldots,d\}\setminus\{k^{(i)}\}}\left[1+\left(\overline{R}_k/(\kappa\underline{R}_k)\right)^{1/\alpha_k}\right]. \end{aligned}$$

Since $\{1,\ldots,L_i\} \subset \mathcal{L}\cup\mathcal{L}'$, we obtain

$$\sum_{j=1}^d |I_j| \sup_{i\in I_j} L_i \leq 4\left[\sum_{j=1}^d\left(1 \vee \log\left(\frac{M_j-m_j}{\eta_j}\right)\right)\right]\left[\prod_{j=1}^d\left(1+\left(\overline{R}_j/(\kappa\underline{R}_j)\right)^{1/\alpha_j}\right)\right],$$

which completes the proof.

7.5.3. Proof of Lemma 2. Without lost of generality and for the sake of simplicity, we assume that $k^{(i)} = d$ and $\psi_d(j) = j$ for all $j \in \{1, \dots, d-1\}$. Let $\ell_1 < \cdots < \ell_r$ be the elements of $\mathcal{L}$. Define for all $p \in \{1, \dots, d-1\}$, $k_{p,0} = 0$ and by induction for all integer $\mathfrak{m}$,

$$k_{p,\mathfrak{m}+1} = \begin{cases} \inf\left\{k > k_{p,\mathfrak{m}},\ \mathrm{j}_{\min}^{(i,\ell_k)} > p\right\} & \text{if there exists } k \in \{k_{p,\mathfrak{m}}+1, \dots, r\} \text{ such that } \mathrm{j}_{\min}^{(i,\ell_k)} > p \\ r & \text{otherwise.} \end{cases}$$

Let $\mathfrak{M}_p$ be the smallest integer $\mathfrak{m}$ for which $k_{p,\mathfrak{m}} = r$. Set for all $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$,

$$K_{p,\mathfrak{m}} = \{k_{p,\mathfrak{m}} + 1, \dots, k_{p,\mathfrak{m}+1}\}.$$

The cardinality of $K_{p,\mathfrak{m}}$ can be upper bounded by the claim below.

Claim 3. *For all $p \in \{1, \dots, d-1\}$ and $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$,*

$$|K_{p,\mathfrak{m}}| \le \prod_{k=1}^{p} \left[1 + \left(\frac{\overline{R}_k}{\kappa \underline{R}_k}\right)^{1/\alpha_k}\right]. \tag{27}$$

Lemma 2 follows from the equality $\mathcal{L} = K_{d-1,0}$. The cardinality of $\mathcal{L}'$ can be bounded from above in the same way. □

Proof of Claim 3. The result is proved by induction. We begin to prove (27) when $p = 1$.

Let $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_1 - 1\}$. We have $\theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+1})} = a_1^{(i)}$ and for $j \in \{1, \dots, k_{1,\mathfrak{m}+1} - k_{1,\mathfrak{m}} - 1\}$,

$$\theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+j+1})} \ge \theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+j})} + \bar{r}_{\Theta_i,1}\left(\boldsymbol{\theta}^{(i,\ell_{k_{1,\mathfrak{m}}+j})}, \boldsymbol{\theta}'^{(i,\ell_{k_{1,\mathfrak{m}}+j})}\right).$$

Now,

$$\bar{r}_{\Theta_i,1}\left(\boldsymbol{\theta}^{(i,\ell_{k_{1,\mathfrak{m}}+j})}, \boldsymbol{\theta}'^{(i,\ell_{k_{1,\mathfrak{m}}+j})}\right) \ge \left((\kappa \underline{R}_{\Theta_i,d}/\overline{R}_{\Theta_i,1})(b_d^{(i)} - a_d^{(i)})^{\alpha_d}\right)^{1/\alpha_1}.$$

Since $\underline{R}_{\Theta_i,d}(b_d^{(i)} - a_d^{(i)})^{\alpha_d} \ge \underline{R}_{\Theta_i,1}(b_1^{(i)} - a_1^{(i)})^{\alpha_1}$,

$$\begin{aligned} \bar{r}_{\Theta_i,1}\left(\boldsymbol{\theta}^{(i,\ell_{k_{1,\mathfrak{m}}+j})}, \boldsymbol{\theta}'^{(i,\ell_{k_{1,\mathfrak{m}}+j})}\right) &\ge \left(\kappa \underline{R}_{\Theta_i,1}/\overline{R}_{\Theta_i,1}\right)^{1/\alpha_1}(b_1^{(i)} - a_1^{(i)}) \\ &\ge (\kappa \underline{R}_1/\overline{R}_1)^{1/\alpha_1}(b_1^{(i)} - a_1^{(i)}). \end{aligned} \tag{28}$$

This leads to

$$\theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+j+1})} \ge \theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+j})} + \left(\kappa \underline{R}_1/\overline{R}_1\right)^{1/\alpha_1}(b_1^{(i)} - a_1^{(i)}).$$

Moreover, $\theta_1^{(i,\ell_{k_{1,\mathfrak{m}+1}})} \le b_1^{(i)}$ (because all the $\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}$ belong to Θ_i). Consequently,

$$a_1^{(i)} + (k_{1,\mathfrak{m}+1} - k_{1,\mathfrak{m}} - 1)\left(\kappa \underline{R}_1/\overline{R}_1\right)^{1/\alpha_1}\left(b_1^{(i)} - a_1^{(i)}\right) \le b_1^{(i)},$$

which shows the result for $p = 1$.

Suppose now that (27) holds for $p \in \{1, \dots, d-2\}$. We shall show that it also holds for $p+1$. Let $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_{p+1} - 1\}$. We use the claim below whose proof is postponed to Section 7.5.6.

Claim 4. *For all* $\mathfrak{m} \in \{0, \ldots, \mathfrak{M}_{p+1} - 1\}$*, there exists* $\mathfrak{m}' \in \{0, \ldots, \mathfrak{M}_p - 1\}$ *such that* $k_{p,\mathfrak{m}'+1} \in K_{p+1,\mathfrak{m}}$.

The claim says that we can consider the smallest integer $\mathfrak{m}_0$ of $\{0, \ldots, \mathfrak{M}_p - 1\}$ such that $k_{p,\mathfrak{m}_0+1} > k_{p+1,\mathfrak{m}}$, and the larger integer $\mathfrak{m}_1$ of $\{0, \ldots, \mathfrak{M}_p - 1\}$ such that $k_{p,\mathfrak{m}_1+1} \leq k_{p+1,\mathfrak{m}+1}$. We define

$$\begin{aligned} I_{\mathfrak{m}_0} &= \{k_{p+1,\mathfrak{m}} + 1, \ldots, k_{p,\mathfrak{m}_0+1}\} \\ I_{\mathfrak{m}'} &= \left\{k_{p,\mathfrak{m}'} + 1, \ldots, k_{p,\mathfrak{m}'+1}\right\} \quad \text{for all } \mathfrak{m}' \in \{\mathfrak{m}_0 + 1, \ldots, \mathfrak{m}_1\} \\ I_{\mathfrak{m}_1+1} &= \{k_{p,\mathfrak{m}_1+1} + 1, \ldots, k_{p+1,\mathfrak{m}+1}\}. \end{aligned}$$

We then have

$$K_{p+1,\mathfrak{m}} = \bigcup_{\mathfrak{m}'=\mathfrak{m}_0}^{\mathfrak{m}_1+1} I_{\mathfrak{m}'}.$$

Notice that for all $\mathfrak{m}' \in \{\mathfrak{m}_0, \ldots, \mathfrak{m}_1\}$, $I_{\mathfrak{m}'} \subset K_{p,\mathfrak{m}'}$. We consider two cases.

- If $k_{p,\mathfrak{m}_1+1} = k_{p+1,\mathfrak{m}+1}$, then $I_{\mathfrak{m}_1+1} = \emptyset$ and thus, by using the above inclusion and the induction assumption,

$$|K_{p+1,\mathfrak{m}'}| \leq (\mathfrak{m}_1 - \mathfrak{m}_0 + 1) \prod_{k=1}^{p} \left[1 + \left(\frac{\overline{R}_k}{\kappa \underline{R}_k}\right)^{1/\alpha_k}\right].$$

- If $k_{p,\mathfrak{m}_1+1} < k_{p+1,\mathfrak{m}+1}$ then $\mathfrak{m}_1 + 1 \leq \mathfrak{M}_p - 1$. Indeed, if this is not true, then $\mathfrak{m}_1 = \mathfrak{M}_p - 1$, which leads to $k_{p,\mathfrak{m}_1+1} = r$ and thus $k_{p+1,\mathfrak{m}+1} > r$. This is impossible since $k_{p+1,\mathfrak{m}+1}$ is always smaller than r (by definition). Consequently, $I_{\mathfrak{m}_1+1} \subset K_{p,\mathfrak{m}_1+1}$ and we derive from the induction assumption,

$$|K_{p+1,\mathfrak{m}'}| \leq (\mathfrak{m}_1 - \mathfrak{m}_0 + 2) \prod_{k=1}^{p} \left[1 + \left(\frac{\overline{R}_k}{\kappa \underline{R}_k}\right)^{1/\alpha_k}\right].$$

We now bound from above $\mathfrak{m}_1 - \mathfrak{m}_0$.

Since for all $k \in \{k_{p+1,\mathfrak{m}} + 1, \ldots, k_{p,\mathfrak{m}_0+1} - 1\}$, $\mathrm{j}_{\min}^{(i,\ell_k)} \leq p$, we have

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = \theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}}+1})} = a_{p+1}^{(i)}.$$

Since $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = p + 1$,

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}+1})} \geq \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} + \bar{r}_{\Theta_i,p+1}\left(\boldsymbol{\theta}^{(i,1)}, \boldsymbol{\theta}'^{(i,1)}\right)$$

and thus by using a similar argument as the one used in the proof of (28),

$$\begin{aligned} \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}+1})} &\geq \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} + \left(\kappa \underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}} \left(b_{p+1}^{(i)} - a_{p+1}^{(i)}\right) \\ &\geq a_{p+1}^{(i)} + \left(\kappa \underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}} \left(b_{p+1}^{(i)} - a_{p+1}^{(i)}\right). \end{aligned}$$

Similarly, for all $\mathfrak{m}' \in \{\mathfrak{m}_0+1,\dots,\mathfrak{m}_1\}$ and $k \in \left\{k_{p,\mathfrak{m}'}+1,\dots,k_{p,\mathfrak{m}'+1}-1\right\}$, $\mathrm{j}_{\min}^{(i,\ell_k)} \leq p$ and thus

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}'+1}})} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}'+1}})}.$$

Moreover, for all $\mathfrak{m}' \in \{\mathfrak{m}_0+1,\dots,\mathfrak{m}_1-1\}$, $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}'+1}})} = p+1$ and thus

$$\begin{aligned}\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}'+1}+1})} &\geq \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}'+1}})} + \left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right) \\ &\geq \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}'+1}})} + \left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right).\end{aligned} \tag{29}$$

This leads to

$$\begin{aligned}\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1}+1})} &\geq \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}+1})} + (\mathfrak{m}_1-\mathfrak{m}_0-1)\left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right) \\ &\geq a_{p+1}^{(i)} + (\mathfrak{m}_1-\mathfrak{m}_0)\left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right).\end{aligned}$$

There are two types of cases involved: if $k_{p,\mathfrak{m}_1+1} = k_{p+1,\mathfrak{m}+1}$ and if $k_{p,\mathfrak{m}_1+1} < k_{p+1,\mathfrak{m}+1}$.

- If $k_{p,\mathfrak{m}_1+1} = k_{p+1,\mathfrak{m}+1}$,

  $$\begin{aligned}\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}+1}})} &= \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1}+1})} \\ &\geq a_{p+1}^{(i)} + (\mathfrak{m}_1-\mathfrak{m}_0)\left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right).\end{aligned}$$

  Since $\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}+1}})} \leq b_{p+1}^{(i)}$, we have

  $$\mathfrak{m}_1-\mathfrak{m}_0 \leq \left(\overline{R}_{p+1}/(\kappa\underline{R}_{p+1})\right)^{1/\alpha_{p+1}}.$$

- If now $k_{p,\mathfrak{m}_1+1} < k_{p+1,\mathfrak{m}+1}$, then (29) also holds for $\mathfrak{m}' = \mathfrak{m}_1$. This implies

  $$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}+1})} \geq a_{p+1}^{(i)} + (\mathfrak{m}_1-\mathfrak{m}_0+1)\left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right).$$

  Since $\mathrm{j}_{\min}^{(i,\ell_k)} \leq p$ for all $k \in \{k_{p,\mathfrak{m}_1+1},\dots,k_{p+1,\mathfrak{m}+1}-1\}$,

  $$\begin{aligned}\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}+1}})} &= \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}+1})} \\ &\geq a_{p+1}^{(i)} + (\mathfrak{m}_1-\mathfrak{m}_0+1)\left(\kappa\underline{R}_{p+1}/\overline{R}_{p+1}\right)^{1/\alpha_{p+1}}\left(b_{p+1}^{(i)}-a_{p+1}^{(i)}\right).\end{aligned}$$

  Since, $\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}+1}})} \leq b_{p+1}^{(i)}$,

  $$\mathfrak{m}_1-\mathfrak{m}_0+1 \leq \left(\overline{R}_{p+1}/(\kappa\underline{R}_{p+1})\right)^{1/\alpha_{p+1}}.$$

This ends the proof. □

7.5.4. Proof of Theorem 5. The lemma and claim below show that the assumptions of Theorem 7 (page 134) are satisfied.

Lemma 3. *For all $i \in \{1, \dots, N-1\}$,*

$$\Theta_i \setminus \bigcup_{\ell=1}^{L_i} B^{(i,\ell)} \subset \Theta_{i+1} \subset \Theta_i.$$

Claim 5. *For all $i \in \{1, \dots, N-1\}$ and $\ell \in \{1, \dots, L_i\}$,*

$$\kappa_0 \, diam(\Theta_i) \leq h^2(f_{\boldsymbol{\theta}^{(i,\ell)}}, f_{\boldsymbol{\theta}'^{(i,\ell)}})$$

where $\kappa_0 = \inf_{1 \leq j \leq d} \underline{R}_j / \overline{R}_j$.

We now derive from Theorem 7 that

$$\mathbb{P}\left[C \inf_{\boldsymbol{\theta} \in \Theta_N} h^2(s, f_{\boldsymbol{\theta}}) \geq h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi \right] \leq e^{-n\xi}$$

where $C > 0$ depends only on $\kappa, \sup_{1 \leq j \leq d} \overline{R}_j / \underline{R}_j$. Consequently, with probability larger than $1 - e^{-n\xi}$,

$$\begin{aligned} h^2(s, f_{\hat{\boldsymbol{\theta}}}) &\leq 2 \inf_{\boldsymbol{\theta} \in \Theta_N} h^2(s, f_{\boldsymbol{\theta}}) + 2\mathrm{diam}\, \Theta_N \\ &\leq 2C^{-1} \left(h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi \right) + 2 \sup_{1 \leq j \leq d} \overline{R}_j \eta_j^{\alpha_j} \\ &\leq 2C^{-1} \left(h^2(s, \mathscr{F}) + \frac{D_{\mathscr{F}}}{n} + \xi \right) + 2d/n \\ &\leq C' \left(h^2(s, \mathscr{F}) + \frac{d}{n} + \xi \right). \end{aligned}$$

The theorem follows. □

7.5.5. Proof of Lemma 3. Since

$$\varepsilon_{k^{(i)}}^{(i,L_i)} \leq \frac{b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)}}{2} \quad \text{and} \quad \varepsilon_{k^{(i)}}^{\prime(i,L_i)} \leq \frac{b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)}}{2},$$

we have $\Theta_{i+1} \subset \Theta_i$. We now aim at proving $\Theta_i \setminus \cup_{\ell=1}^{L_i} B^{(i,\ell)} \subset \Theta_{i+1}$.

We introduce the rectangles

$$\begin{aligned} \mathcal{R}^{(i,\ell)} &= \prod_{q=1}^{d} \left[\theta_q^{(i,\ell)}, \theta_q^{(i,\ell)} + \varepsilon_q^{(i,\ell)} \right] \\ \mathcal{R}'^{(i,\ell)} &= \prod_{q=1}^{k^{(i)}-1} \left[\theta_q'^{(i,\ell)}, \theta_q'^{(i,\ell)} + \varepsilon_q'^{(i,\ell)} \right] \times \left[\theta_{k^{(i)}}'^{(i,\ell)} - \varepsilon_{k^{(i)}}'^{(i,\ell)}, \theta_{k^{(i)}}'^{(i,\ell)} \right] \times \prod_{q=k^{(i)}+1}^{d} \left[\theta_q'^{(i,\ell)}, \theta_q'^{(i,\ell)} + \varepsilon_q'^{(i,\ell)} \right] \end{aligned}$$

and we set

$$\mathcal{R}''^{(i,\ell)} = \begin{cases} \mathcal{R}^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) > 0 \\ \mathcal{R}'^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) < 0 \\ \mathcal{R}^{(i,\ell)} \bigcup \mathcal{R}'^{(i,\ell)} & \text{if } T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) = 0. \end{cases}$$

We derive from (15) that $\Theta_i \cap \mathcal{R}''^{(i,\ell)} \subset B^{(i,\ell)}$. It is then sufficient to show

$$\Theta_i \setminus \bigcup_{\ell=1}^{L_i} \mathcal{R}''^{(i,\ell)} \subset \Theta_{i+1}.$$

For this purpose, note that either $T(\boldsymbol{\theta}^{(i,L_i)}, \boldsymbol{\theta}'^{(i,L_i)}) \geq 0$ or $T(\boldsymbol{\theta}^{(i,L_i)}, \boldsymbol{\theta}'^{(i,L_i)}) \leq 0$. In what follows, we assume that $T(\boldsymbol{\theta}^{(i,L_i)}, \boldsymbol{\theta}'^{(i,L_i)}) \geq 0$ but a similar proof can be made if $T(\boldsymbol{\theta}^{(i,L_i)}, \boldsymbol{\theta}'^{(i,L_i)})$ is non-positive. Without lost of generality, and for the sake of simplicity, we suppose as in the proof of Lemma 2 that $k^{(i)} = d$ and $\psi_d(j) = j$ for all $j \in \{1, \ldots, d-1\}$. Let

$$\mathcal{L} = \left\{ 1 \leq \ell \leq L_i,\ T(\boldsymbol{\theta}^{(i,\ell)}, \boldsymbol{\theta}'^{(i,\ell)}) \geq 0 \right\}$$

and $\ell_1 < \cdots < \ell_r$ be the elements of $\mathcal{L}$. We have

$$\Theta_{i+1} = \prod_{q=1}^{d-1} \left[a_q^{(i)}, b_q^{(i)} \right] \times \left[a_d^{(i)} + \varepsilon_d^{(i,L_i)}, b_d^{(i)} \right]$$

and it is sufficient to prove that

$$\prod_{q=1}^{d-1} \left[a_q^{(i)}, b_q^{(i)} \right] \times \left[a_d^{(i)}, a_d^{(i)} + \varepsilon_d^{(i,L_i)} \right] \subset \bigcup_{k=1}^{r} \mathcal{R}^{(i,\ell_k)}.$$

For this, remark that for all $k \in \{1, \ldots, r\}$, $\theta_d^{(i,\ell_k)} = a_d^{(i)}$ and thus

$$\mathcal{R}^{(i,\ell_k)} = \prod_{q=1}^{d-1} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)} \right] \times \left[a_d^{(i)}, a_d^{(i)} + \varepsilon_d^{(i,\ell_k)} \right].$$

By using the fact that the sequence $(\varepsilon_d^{(i,\ell_k)})_k$ is non-increasing,

$$\left[a_d^{(i)}, a_d^{(i)} + \varepsilon_d^{(i,L_i)} \right] \subset \bigcap_{k=1}^{r} \left[a_d^{(i)}, a_d^{(i)} + \varepsilon_d^{(i,\ell_k)} \right].$$

This means that it is sufficient to show that

$$\prod_{q=1}^{d-1} \left[a_q^{(i)}, b_q^{(i)} \right] \subset \bigcup_{k=1}^{r} \prod_{q=1}^{d-1} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)} \right]. \tag{30}$$

Let us now define (as in the proof of Lemma 2) for all $p \in \{1, \ldots, d-1\}$, $k_{p,0} = 0$ and by induction for all integer $\mathfrak{m}$,

$$k_{p,\mathfrak{m}+1} = \begin{cases} \inf \left\{ k > k_{p,\mathfrak{m}},\ \mathrm{j}_{\min}^{(i,\ell_k)} > p \right\} & \text{if there exists } k \in \{k_{p,\mathfrak{m}}+1, \ldots, r\} \text{ such that } \mathrm{j}_{\min}^{(i,\ell_k)} > p \\ r & \text{otherwise.} \end{cases}$$

Let $\mathfrak{M}_p$ be the smallest integer $\mathfrak{m}$ such that $k_{p,\mathfrak{m}} = r$. Let then for all $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$,

$$K_{p,\mathfrak{m}} = \{k_{p,\mathfrak{m}} + 1, \dots, k_{p,\mathfrak{m}+1}\}.$$

We shall use the claim below (whose proof is delayed to Section 7.5.6).

Claim 6. *Let* $\mathfrak{m}' \in \{0, \dots, \mathfrak{M}_{p+1}-1\}$, $p \in \{1, \dots, d-1\}$. *There exists a subset* $\mathcal{M}$ *of* $\{0, \dots, \mathfrak{M}_p - 1\}$ *such that*

$$K'_p = \{k_{p,\mathfrak{m}+1},\ \mathfrak{m} \in \mathcal{M}\} \subset K_{p+1,\mathfrak{m}'}$$

and

$$\left[a_{p+1}^{(i)}, b_{p+1}^{(i)}\right] \subset \bigcup_{k \in K'_p} \left[\theta_{p+1}^{(i,\ell_k)}, \theta_{p+1}^{(i,\ell_k)} + \varepsilon_{p+1}^{(i,\ell_k)}\right].$$

We prove by induction on p the following result. For all $p \in \{1, \dots, d-1\}$, and all $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$,

$$\prod_{q=1}^{p} \left[a_q^{(i)}, b_q^{(i)}\right] \subset \bigcup_{k \in K_{p,\mathfrak{m}}} \prod_{q=1}^{p} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)}\right] \tag{31}$$

Note that (30) ensues from this inclusion when $p = d-1$ and $\mathfrak{m} = 0$.

We begin to prove (31) for $p = 1$ and all $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_1 - 1\}$. For all $k \in \{k_{1,\mathfrak{m}}+1, \dots, k_{1,\mathfrak{m}+1}-1\}$, $\mathrm{j}_{\min}^{(i,\ell_k)} \leq 1$ and thus

$$\theta_1^{(i,\ell_{k+1})} \in \left\{\theta_1^{(i,\ell_k)}, \theta_1^{(i,\ell_k)} + \varepsilon_1^{(i,\ell_k)}\right\}.$$

This implies that the set

$$\bigcup_{k=k_{1,\mathfrak{m}}+1}^{k_{1,\mathfrak{m}+1}} \left[\theta_1^{(i,\ell_k)}, \theta_1^{(i,\ell_k)} + \varepsilon_1^{(i,\ell_k)}\right].$$

is an interval. Now, $\theta_1^{(i,\ell_{k_{1,\mathfrak{m}}+1})} = a_1^{(i)}$, $\theta_1^{(i,\ell_{k_{1,\mathfrak{m}+1}})} + \varepsilon_1^{(i,\ell_{k_{1,\mathfrak{m}+1}})} \geq b_1^{(i)}$ since $\mathrm{j}_{\min}^{(i,\ell_{k_{1,\mathfrak{m}+1}})} > 1$. We have

$$[a_1^{(i)}, b_1^{(i)}] \subset \bigcup_{k=k_{1,\mathfrak{m}}+1}^{k_{1,\mathfrak{m}+1}} \left[\theta_1^{(i,\ell_k)}, \theta_1^{(i,\ell_k)} + \varepsilon_1^{(i,\ell_k)}\right]$$

which establishes (31) when $p = 1$.

Let now $p \in \{1, \dots, d-2\}$ and assume that for all $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$,

$$\prod_{q=1}^{p} \left[a_q^{(i)}, b_q^{(i)}\right] \subset \bigcup_{k \in K_{p,\mathfrak{m}}} \prod_{q=1}^{p} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)}\right].$$

Let $\mathfrak{m}' \in \{0, \dots, \mathfrak{M}_{p+1} - 1\}$. We shall show that

$$\prod_{q=1}^{p+1} \left[a_q^{(i)}, b_q^{(i)}\right] \subset \bigcup_{k \in K_{p+1,\mathfrak{m}'}} \prod_{q=1}^{p+1} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)}\right].$$

Let $\boldsymbol{x} \in \prod_{q=1}^{p+1} \left[a_q^{(i)}, b_q^{(i)}\right]$. By using Claim 6, there exists $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$ such that

$$x_{p+1} \in \left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}\right]$$

and such that $k_{p,\mathfrak{m}+1} \in K_{p+1,\mathfrak{m}'}$. By using the induction assumption, there exists $k \in K_{p,\mathfrak{m}}$ such that

$$\boldsymbol{x} = (x_1, \dots, x_p) \in \prod_{q=1}^{p} \left[\theta_q^{(i,\ell_k)}, \theta_q^{(i,\ell_k)} + \varepsilon_q^{(i,\ell_k)}\right].$$

Since $k \in K_{p,\mathfrak{m}}$, $\theta_{p+1}^{(i,\ell_k)} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}$ and $\varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})} \leq \varepsilon_{p+1}^{(i,\ell_k)}$. Hence,

$$x_{p+1} \in \left[\theta_{p+1}^{(i,\ell_k)}, \theta_{p+1}^{(i,\ell_k)} + \varepsilon_{p+1}^{(i,\ell_k)}\right].$$

We finally use the claim below to show that $k \in K_{p+1,\mathfrak{m}'}$ which concludes the proof.

Claim 7. *Let $\mathfrak{m} \in \{0, \dots, \mathfrak{M}_p - 1\}$ and $\mathfrak{m}' \in \{0, \dots, \mathfrak{M}_{p+1} - 1\}$. If $k_{p,\mathfrak{m}+1} \in K_{p+1,\mathfrak{m}'}$, then $K_{p,\mathfrak{m}} \subset K_{p+1,\mathfrak{m}'}$.*

7.5.6. Proof of the claims.

Proof of Claim 4. The set $\{\mathfrak{m}' \in \{0, \dots, \mathfrak{M}_p - 1\},\ k_{p,\mathfrak{m}'+1} \leq k_{p+1,\mathfrak{m}+1}\}$ is non empty and we can thus define the largest integer $\mathfrak{m}'$ of $\{0, \dots, \mathfrak{M}_p - 1\}$ such that $k_{p,\mathfrak{m}'+1} \leq k_{p+1,\mathfrak{m}+1}$. We then have

$$k_{p,\mathfrak{m}'} = \sup\left\{k < k_{p,\mathfrak{m}'+1},\ \mathfrak{j}_{\min}^{(i,\ell_k)} > p\right\}.$$

Since $k_{p,\mathfrak{m}'} < k_{p+1,\mathfrak{m}+1}$,

$$\begin{aligned} k_{p,\mathfrak{m}'} &= \sup\left\{k < k_{p+1,\mathfrak{m}+1},\ \mathfrak{j}_{\min}^{(i,\ell_k)} > p\right\} \\ &\geq \sup\left\{k < k_{p+1,\mathfrak{m}+1},\ \mathfrak{j}_{\min}^{(i,\ell_k)} > p+1\right\} \\ &\geq k_{p+1,\mathfrak{m}}. \end{aligned}$$

Since, $k_{p,\mathfrak{m}'+1} \geq k_{p,\mathfrak{m}'} + 1$, we also have $k_{p,\mathfrak{m}'+1} \geq k_{p+1,\mathfrak{m}} + 1$. Finally, $k_{p,\mathfrak{m}'+1} \in K_{p,\mathfrak{m}}$. □

Proof of Claim 5. Let $i \in \{1, \dots, N-1\}$ and $\ell \in \{1, \dots, L_i\}$. Then,

$$\begin{aligned} \operatorname{diam}(\Theta_i) &\leq \sup_{1\leq j\leq d} \overline{R}_{\Theta_i,j}\big(b_j^{(i)} - a_j^{(i)}\big)^{\alpha_j} \\ &\leq \left(\sup_{1\leq j\leq d} \frac{\overline{R}_{\Theta_i,j}}{\underline{R}_{\Theta_i,j}}\right) \sup_{1\leq j\leq d} \underline{R}_{\Theta_i,j}\big(b_j^{(i)} - a_j^{(i)}\big)^{\alpha_j} \\ &\leq \left(\sup_{1\leq j\leq d} \frac{\overline{R}_{\Theta_i,j}}{\underline{R}_{\Theta_i,j}}\right) \underline{R}_{\Theta_i,k^{(i)}}\big(b_{k^{(i)}}^{(i)} - a_{k^{(i)}}^{(i)}\big)^{\alpha_{k^{(i)}}}. \end{aligned}$$

Now, $\theta_{k^{(i)}}^{(i,\ell)} = a_{k^{(i)}}^{(i)}$ and $\theta_{k^{(i)}}^{\prime(i,\ell)} = b_{k^{(i)}}^{(i)}$ and thus

$$\begin{aligned}
\operatorname{diam}(\Theta_i) &\leq \left(\sup_{1\leq j\leq d} \frac{\overline{R}_{\Theta_i,j}}{\underline{R}_{\Theta_i,j}} \right) \underline{R}_{\Theta_i,k^{(i)}} \big(\theta_{k^{(i)}}^{(i,\ell)} - \theta_{k^{(i)}}^{(i,\ell)}\big)^{\alpha_{k^{(i)}}} \\
&\leq \left(\sup_{1\leq j\leq d} \frac{\overline{R}_{\Theta_i,j}}{\underline{R}_{\Theta_i,j}} \right) \sup_{1\leq j\leq d} \underline{R}_{\Theta_i,j} \big(\theta_j^{\prime(i,\ell)} - \theta_j^{(i,\ell)}\big)^{\alpha_j} \\
&\leq \left(\sup_{1\leq j\leq d} \frac{\overline{R}_{\Theta_i,j}}{\underline{R}_{\Theta_i,j}} \right) h^2(f_{\boldsymbol{\theta}^{(i,\ell)}}, f_{\boldsymbol{\theta}^{\prime(i,\ell)}}).
\end{aligned}$$

We conclude by using $\overline{R}_{\Theta_i,j}/\underline{R}_{\Theta_i,j} \leq \overline{R}_j/\underline{R}_j$. □

Proof of Claim 6. Thanks to Claim 4 (page 146), we can define the smallest integer $\mathfrak{m}_0$ of $\{0,\ldots,\mathfrak{M}_p - 1\}$ such that $k_{p,\mathfrak{m}_0+1} \in K_{p+1,\mathfrak{m}'}$, and the largest integer $\mathfrak{m}_1$ of $\{0,\ldots,\mathfrak{M}_p - 1\}$ such that $k_{p,\mathfrak{m}_1+1} \in K_{p+1,\mathfrak{m}'}$. Define now

$$\mathcal{M} = \{\mathfrak{m}_0, \mathfrak{m}_0 + 1, \ldots, \mathfrak{m}_1\}.$$

Note that for all $\mathfrak{m} \in \{\mathfrak{m}_0,\ldots,\mathfrak{m}_1\}$, $k_{p,\mathfrak{m}+1} \in K_{p+1,\mathfrak{m}'}$ (this ensues from the fact that the sequence $(k_{p,\mathfrak{m}})_{\mathfrak{m}}$ is increasing).

Let $\mathfrak{m} \in \{0,\ldots,\mathfrak{M}_p - 1\}$ be such that $k_{p,\mathfrak{m}} \in K_{p+1,\mathfrak{m}'}$ and $k_{p,\mathfrak{m}} \neq k_{p+1,\mathfrak{m}'+1}$. Then $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}}})} \leq p+1$ and since $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}}})} > p$, we also have $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}}})} = p+1$. Consequently,

$$\theta_{p+1}^{\left(i,\ell_{k_{p,\mathfrak{m}+1}}\right)} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})}.$$

Now, $\theta_{p+1}^{\left(i,\ell_{k_{p,\mathfrak{m}+1}}\right)} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}$ since $k_{p,\mathfrak{m}}+1$ and $k_{p,\mathfrak{m}+1}$ belong to $K_{p,\mathfrak{m}}$. The set

$$\left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})}\right] \bigcup \left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}\right]$$

is thus the interval

$$\left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}\right].$$

We apply this argument for each $\mathfrak{m} \in \{\mathfrak{m}_0+1,\ldots,\mathfrak{m}_1\}$ to derive that the set

$$I = \bigcup_{\mathfrak{m}=\mathfrak{m}_0}^{\mathfrak{m}_1} \left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}+1}})}\right]$$

is the interval

$$I = \left[\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})}, \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})}\right].$$

The claim is proved if we show that

$$\left[a_{p+1}^{(i)}, b_{p+1}^{(i)}\right] \subset I.$$

Since I is an interval, it remains to prove that $a_{p+1}^{(i)} \in I$ and $b_{p+1}^{(i)} \in I$.

We begin to show $a_{p+1}^{(i)} \in I$ by showing that $a_{p+1}^{(i)} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})}$. If $k_{p+1,\mathfrak{m}'} = 0$, then $\mathfrak{m}' = 0$ and $\mathfrak{m}_0 = 0$. Besides, since 1 and $k_{p,1}$ belong to $K_{p,0}$, we have $\theta_{p+1}^{(i,\ell_{k_{p,1}})} = \theta_{p+1}^{(i,\ell_1)}$. Now, $\theta_{p+1}^{(i,\ell_1)} = a_{p+1}^{(i)}$ and thus $a_{p+1}^{(i)} \in I$. We now assume that $k_{p+1,\mathfrak{m}'} \neq 0$. Since $k_{p,\mathfrak{m}_0} \leq k_{p+1,\mathfrak{m}'}$, there are two cases.

- First case: $k_{p,\mathfrak{m}_0} = k_{p+1,\mathfrak{m}'}$. We then have $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}_0}})} > p+1$ and thus $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = a_{p+1}^{(i)}$. Since $k_{p,\mathfrak{m}_0+1}$ and $k_{p,\mathfrak{m}_0}+1$ belong to $K_{p,\mathfrak{m}_0}$, $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})}$ and thus $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = a_{p+1}^{(i)}$ as wished.
- Second case: $k_{p,\mathfrak{m}_0} + 1 \leq k_{p+1,\mathfrak{m}'}$. Then, $k_{p+1,\mathfrak{m}'} \in K_{p,\mathfrak{m}_0}$ and thus

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = \theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}'}})}.$$

  Since $\mathrm{j}_{\min}^{(i,\ell_{k_{p+1,\mathfrak{m}'}})} > p+1$, we have $\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}'}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}'}})} \geq b_{p+1}^{(i)}$. By using the fact that the sequence $(\varepsilon_{p+1}^{(i,\ell_k)})_k$ is decreasing, we then deduce

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} \geq b_{p+1}^{(i)}$$

  and thus $\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} > p+1$. This establishes that

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+2}})} = a_{p+1}^{(i)}. \tag{32}$$

  Let us now show $k_{p,\mathfrak{m}_0} + 2 \leq k_{p,\mathfrak{m}_0+1}$. Otherwise, $k_{p,\mathfrak{m}_0} + 2 \geq k_{p,\mathfrak{m}_0+1} + 1$ and thus $k_{p,\mathfrak{m}_0} + 1 \geq k_{p,\mathfrak{m}_0+1}$ which means that $k_{p,\mathfrak{m}_0} + 1 = k_{p,\mathfrak{m}_0+1}$ (we recall that $(k_{p,\mathfrak{m}})_{\mathfrak{m}}$ is an increasing sequence of integers). Since we are in the case where $k_{p,\mathfrak{m}_0} + 1 \leq k_{p+1,\mathfrak{m}'}$, we have $k_{p,\mathfrak{m}_0+1} \leq k_{p+1,\mathfrak{m}'}$ which is impossible since $k_{p,\mathfrak{m}_0+1} \in K_{p+1,\mathfrak{m}'}$.

  Consequently, since $k_{p,\mathfrak{m}_0} + 2 \leq k_{p,\mathfrak{m}_0+1}$, we have $k_{p,\mathfrak{m}_0} + 2 \in K_{p,\mathfrak{m}_0}$ and thus $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = \theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+2}})}$. We then deduce from (32) that $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_0+1}})} = a_{p+1}^{(i)}$ as wished.

We now show that $b_{p+1}^{(i)} \in I$ by showing that $\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})} \geq b_{p+1}^{(i)}$. If $\mathfrak{m}_1 = \mathfrak{M}_p - 1$,

$$\theta_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})} = \theta_{p+1}^{(i,\ell_r)} + \varepsilon_{p+1}^{(i,\ell_r)} = \theta_{p+1}^{(i,L_i)} + \varepsilon_{p+1}^{(i,L_i)}.$$

Since $\mathrm{j}_{\min}^{(i,L_i)} = d$, we have $\theta_{p+1}^{(i,L_i)} + \varepsilon_{p+1}^{(i,L_i)} \geq b_{p+1}^{(i)}$ which proves the result.

We now assume that $\mathfrak{m}_1 < \mathfrak{M}_p - 1$. We begin to prove that $k_{p,\mathfrak{m}_1+1} = k_{p+1,\mathfrak{m}'+1}$. If this inequality does not hold, we derive from the inequality $k_{p,\mathfrak{m}_1+1} \leq k_{p+1,\mathfrak{m}'+1} < k_{p,\mathfrak{m}_1+2}$, that $k_{p,\mathfrak{m}_1+1} + 1 \leq k_{p+1,\mathfrak{m}'+1}$ and thus $k_{p+1,\mathfrak{m}'+1} \in K_{p,\mathfrak{m}_1+1}$. Since $\mathrm{j}_{\min}^{(i,\ell_{k_{p+1,\mathfrak{m}'+1}})} > p+1$, we have

$$\theta_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}'+1}})} + \varepsilon_{p+1}^{(i,\ell_{k_{p+1,\mathfrak{m}'+1}})} \geq b_{p+1}^{(i)}.$$

Hence,

$$\theta_{p+1}^{(i,\ell_{(k_{p,\mathfrak{m}_1+1})+1})}+\varepsilon_{p+1}^{(i,\ell_{(k_{p,\mathfrak{m}_1+1})+1})}\geq b_{p+1}^{(i)}\quad\text{which implies}\quad \mathrm{j}_{\min}^{(i,\ell_{(k_{p,\mathfrak{m}_1+1})+1})}>p+1.$$

Since,

$$k_{p+1,\mathfrak{m}'+1}=\inf\left\{k>k_{p+1,\mathfrak{m}'},\ \mathrm{j}_{\min}^{(i,\ell_k)}>p+1\right\}$$

and $k_{p,\mathfrak{m}_1+1}+1>k_{p+1,\mathfrak{m}'}$, we have $k_{p+1,\mathfrak{m}'+1}\leq k_{p,\mathfrak{m}_1+1}+1$. Moreover, since $k_{p+1,\mathfrak{m}'+1}\geq k_{p,\mathfrak{m}_1+1}+1$, we have $k_{p,\mathfrak{m}_1+1}+1=k_{p+1,\mathfrak{m}'+1}$. Consequently,

$$k_{p,\mathfrak{m}_1+2}=\inf\left\{k>k_{p,\mathfrak{m}_1+1},\ \mathrm{j}_{\min}^{(i,\ell_k)}>p\right\}=k_{p+1,\mathfrak{m}'+1}.$$

This is impossible because $k_{p+1,\mathfrak{m}'+1}<k_{p,\mathfrak{m}_1+2}$, which finally implies that $k_{p,\mathfrak{m}_1+1}=k_{p+1,\mathfrak{m}'+1}$.

We then deduce from this equality,

$$\mathrm{j}_{\min}^{(i,\ell_{k_{p,\mathfrak{m}_1+1}})}=\mathrm{j}_{\min}^{(i,\ell_{k_{p+1,\mathfrak{m}'+1}})}>p+1.$$

Hence $\theta_{p+1}^{(i,\ell_{\ell_{k_{p,\mathfrak{m}_1+1}}})}+\varepsilon_{p+1}^{(i,\ell_{\ell_{k_{p,\mathfrak{m}_1+1}}})}\geq b_{p+1}^{(i)}$ and thus $b_{p+1}^{(i)}\in I$. This ends the proof. □

Proof of Claim 7. We have

$$k_{p+1,\mathfrak{m}'}=\sup\left\{k<k_{p+1,\mathfrak{m}'+1},\ \mathrm{j}_{\min}^{(i,\ell_k)}>p+1\right\}.$$

Since $k_{p,\mathfrak{m}+1}>k_{p+1,\mathfrak{m}'}$,

$$\begin{aligned}k_{p+1,\mathfrak{m}'} &= \sup\left\{k<k_{p,\mathfrak{m}+1},\ \mathrm{j}_{\min}^{(i,\ell_k)}>p+1\right\}\\ &\leq \sup\left\{k<k_{p,\mathfrak{m}+1},\ \mathrm{j}_{\min}^{(i,\ell_k)}>p\right\}\\ &\leq k_{p,\mathfrak{m}}.\end{aligned}$$

We then derive from the inequalities $k_{p+1,\mathfrak{m}'}\leq k_{p,\mathfrak{m}}$ and $k_{p,\mathfrak{m}+1}\leq k_{p+1'+1}$ that $K_{p,\mathfrak{m}}\subset K_{p+1,\mathfrak{m}'}$.

□

8. Annexe: implementation of the procedure when $d\geq 2$

We carry out in the following sections the values of $\underline{R}_{\mathcal{C},j}$, $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$ and $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$ we have used in the simulation study of Section 6. We do not claim that they minimize the number of tests to compute. The number of tests that have been computed in the simulation study with these choices of parameters may be found in Section 8.7.

8.1. Example 1. In the case of the Gaussian model, it is worthwhile to notice that the Hellinger distance between two densities $f_{(m,\sigma)}$ and $f_{(m',\sigma')}$ can be made explicit:

$$h^2\left(f_{(m,\sigma)}, f_{(m',\sigma')}\right) = 1 - \sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}} \exp\left(-\frac{(m-m')^2}{4(\sigma^2+\sigma'^2)}\right).$$

For all $\xi > 0$, a sufficient condition for that $h^2(f_{(m,\sigma)}, f_{(m',\sigma')}) \leq \xi$ is thus

$$\sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}} \geq \sqrt{1-\xi} \quad \text{and} \quad \exp\left(-\frac{(m-m')^2}{4(\sigma^2+\sigma'^2)}\right) \geq \sqrt{1-\xi}.$$

One then deduces that the rectangle

$$\left[m - 2\frac{1-\sqrt{2\xi-\xi^2}}{1-\xi}\sqrt{\log\left(\frac{1}{1-\xi}\right)}\sigma, m + 2\frac{1-\sqrt{2\xi-\xi^2}}{1-\xi}\sqrt{\log\left(\frac{1}{1-\xi}\right)}\sigma\right] \times \left[\frac{1-\sqrt{2\xi-\xi^2}}{1-\xi}\sigma, \frac{1+\sqrt{2\xi-\xi^2}}{1-\xi}\sigma\right]$$

is included in the Hellinger ball

$$\left\{(m',\sigma') \in \mathbb{R}\times(0,+\infty),\, h^2(f_{(m,\sigma)}, f_{(m',\sigma')}) \leq \xi\right\}.$$

Given $\boldsymbol{\theta} = (m,\sigma)$, $\boldsymbol{\theta}' = (m',\sigma')$, we can then define $\underline{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$, $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$ by

$$\begin{aligned}
\underline{\boldsymbol{r}}_{\mathcal{C}}(\boldsymbol{\theta},\boldsymbol{\theta}') &= \left(2\frac{1-\sqrt{2\xi-\xi^2}}{1-\xi}\sqrt{\log\left(\frac{1}{1-\xi}\right)}\sigma, \frac{-\xi+\sqrt{2\xi-\xi^2}}{1-\xi}\sigma\right)\\
\bar{\boldsymbol{r}}_{\mathcal{C}}(\boldsymbol{\theta},\boldsymbol{\theta}') &= \left(2\frac{1-\sqrt{2\xi-\xi^2}}{1-\xi}\sqrt{\log\left(\frac{1}{1-\xi}\right)}\sigma, \frac{\xi+\sqrt{2\xi-\xi^2}}{1-\xi}\sigma\right)
\end{aligned}$$

where $\xi = \kappa H^2(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})$.

We now consider a rectangle $\mathcal{C} = [\underline{m}_0, \overline{m}_0] \times [\underline{\sigma}_0, \bar{\sigma}_0]$ of $\mathbb{R}\times(0,+\infty)$ and aim at choosing $\underline{\boldsymbol{R}}_{\mathcal{C}} = \left(\underline{R}_{\mathcal{C},1}, \underline{R}_{\mathcal{C},2}\right)$. For all (m,σ), $(m',\sigma') \in \mathcal{C}$,

$$h^2\left(f_{(m,\sigma)}, f_{(m',\sigma')}\right) = 1 - \sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}} + \sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}}\left[1 - \exp\left(-\frac{(m-m')^2}{4(\sigma^2+\sigma'^2)}\right)\right].$$

Yet,

$$1 - \sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}} = \frac{(\sigma'-\sigma)^2}{\left(\sqrt{\sigma^2+\sigma'^2}+\sqrt{2\sigma\sigma'}\right)\sqrt{\sigma^2+\sigma'^2}} \geq \frac{(\sigma'-\sigma)^2}{4\bar{\sigma}_0^2}$$

and

$$\sqrt{\frac{2\sigma\sigma'}{\sigma^2+\sigma'^2}} \geq \sqrt{\frac{2\bar{\sigma}_0\underline{\sigma}_0}{\bar{\sigma}_0^2+\underline{\sigma}_0^2}}.$$

Moreover,

$$1-\exp\left(-\frac{(m-m')^2}{4(\sigma^2+\sigma'^2)}\right)\geq\frac{1-e^{-(\overline{m}_0-\underline{m}_0)^2/(8\bar{\sigma}_0^2)}}{(\overline{m}_0-\underline{m}_0)^2}(m'-m)^2.$$

In particular, we have proved that

$$h^2\left(f_{(m,\sigma)},f_{(m',\sigma')}\right)\geq\max\left\{\sqrt{\frac{2\bar{\sigma}_0\underline{\sigma}_0}{\bar{\sigma}_0^2+\underline{\sigma}_0^2}}\frac{1-e^{-(\overline{m}_0-\underline{m}_0)^2/(8\bar{\sigma}_0^2)}}{(\overline{m}_0-\underline{m}_0)^2}(m'-m)^2,\frac{1}{4\bar{\sigma}_0^2}(\sigma-\sigma')^2\right\}$$

which means that we can take

$$\underline{\boldsymbol{R}}_{\mathcal{C}}=\left(\sqrt{\frac{2\bar{\sigma}_0\underline{\sigma}_0}{\bar{\sigma}_0^2+\underline{\sigma}_0^2}}\frac{1-e^{-(\overline{m}_0-\underline{m}_0)^2/(8\bar{\sigma}_0^2)}}{(\overline{m}_0-\underline{m}_0)^2},\frac{1}{4\bar{\sigma}_0^2}\right).$$

8.2. Example 2. In the case of the Cauchy model, the Hellinger distance cannot be made explicit. However, we can use Theorem 7.6 of Ibragimov and Has'minskii (1981) (Chapter 1) to show that for all $m\in\mathbb{R}$, $\sigma>0$,

$$h^2\left(f_{(0,1)},f_{(m,1)}\right)\leq m^2/16\quad\text{and}\quad h^2\left(f_{(0,1)},f_{(0,\sigma)}\right)\leq(\log^2\sigma)/16.$$

Now,

$$\begin{aligned}h\left(f_{(m,\sigma)},f_{(m',\sigma')}\right)&\leq h\left(f_{(m,\sigma)},f_{(m',\sigma)}\right)+h\left(f_{(m',\sigma)},f_{(m',\sigma')}\right)\\&\leq h\left(f_{(0,1)},f_{((m'-m)/\sigma,1)}\right)+h\left(f_{(0,1)},f_{(0,\sigma'/\sigma)}\right)\\&\leq\frac{|m'-m|}{4\sigma}+\frac{|\log(\sigma'/\sigma)|}{4}.\end{aligned}$$

For all $\xi>0$, one then deduces that the rectangle

$$\left[m-2\sigma\sqrt{\xi},m+2\sigma\sqrt{\xi}\right]\times\left[\sigma e^{-2\sqrt{\xi}},\sigma e^{2\sqrt{\xi}}\right]$$

is included in the Hellinger ball

$$\left\{(m',\sigma')\in\mathbb{R}\times(0,+\infty),\,h^2(f_{(m,\sigma)},f_{(m',\sigma')})\leq\xi\right\}.$$

This provides the values of $\bar{r}_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$ and $\bar{r}'_{\mathcal{C},j}(\boldsymbol{\theta},\boldsymbol{\theta}')$: given $\mathcal{C}\subset\Theta$, $\boldsymbol{\theta}=(m,\sigma)$, $\boldsymbol{\theta}'=(m',\sigma')\in\mathcal{C}$, we can take

$$\begin{aligned}\underline{\boldsymbol{r}}_{\mathcal{C}}(\boldsymbol{\theta},\boldsymbol{\theta}')&=\left(2\sigma\sqrt{\kappa H^2(f_{\boldsymbol{\theta}},f_{\boldsymbol{\theta}'})},\sigma-\sigma e^{-2\sqrt{\kappa H^2(f_{\boldsymbol{\theta}},f_{\boldsymbol{\theta}'})}}\right)\\\bar{\boldsymbol{r}}_{\mathcal{C}}(\boldsymbol{\theta},\boldsymbol{\theta}')&=\left(2\sigma\sqrt{\kappa H^2(f_{\boldsymbol{\theta}},f_{\boldsymbol{\theta}'})},\sigma e^{2\sqrt{\kappa H^2(f_{\boldsymbol{\theta}},f_{\boldsymbol{\theta}'})}}-\sigma\right).\end{aligned}$$

For all rectangle $\mathcal{C}\subset\mathbb{R}\times(0,+\infty)$ we choose $\underline{R}_{\mathcal{C},1}=\underline{R}_{\mathcal{C},2}$. Notice that this choice allows to find easily the number k that appears at line 1 of Algorithm 2 since then the equation becomes

$$b_k-a_k=\max_{1\leq j\leq 2}(b_j-a_j).$$

8.3. Example 3. Let $\xi > 0$, $a, b > 0$ and $\mathcal{C}$ be the rectangle $\mathcal{C} = [a_1, a_2] \times [b_1, b_2] \subset (0, +\infty)^2$. We aim at finding a rectangle $\mathcal{R}$ containing (a, b) such that

$$\mathcal{C} \cap \mathcal{R} \subset \left\{(a', b') \in (0, +\infty)^2,\, h^2(f_{(a,b)}, f_{(a',b')}) \leq \xi\right\}.$$

For this, notice that for all positive numbers a', b',

$$h^2\left(f_{(a,b)}, f_{(a',b')}\right) \leq 2h^2\left(f_{(a,b)}, f_{(a,b')}\right) + 2h^2\left(f_{(a,b')}, f_{(a',b')}\right).$$

Now,

$$h^2\left(f_{(a,b)}, f_{(a,b')}\right) = 1 - \left(\frac{2\sqrt{bb'}}{b+b'}\right)^a.$$

Let Γ' be the derivative of the Gamma function Γ and ψ be the derivative of the digamma function Γ'/Γ. We derive from Theorem 7.6 of Ibragimov and Has'minskii (1981) that

$$h^2\left(f_{(a,b')}, f_{(a',b')}\right) \leq \frac{(a'-a)^2}{8} \sup_{t \in [\min(a,a'), \max(a,a')]} \psi(t).$$

The function ψ being non-increasing,

$$h^2\left(f_{(a,b')}, f_{(a',b')}\right) \leq \begin{cases} 1/8\psi(a)(a'-a)^2 & \text{if } a' \geq a \\ 1/8\psi(a_1)(a'-a)^2 & \text{if } a' < a. \end{cases}$$

We deduce from the above inequalities that we can take

$$\mathcal{R} = \left[a - \sqrt{\frac{2}{\psi(a_1)}\xi}, a + \sqrt{\frac{2}{\psi(a)}\xi}\right] \times \left[\frac{2 - \xi'^2 - 2\sqrt{1-\xi'^2}}{\xi'^2} b, \frac{2 - \xi'^2 + 2\sqrt{1-\xi'^2}}{\xi'^2} b\right]$$

where $\xi' = (1 - \xi/4)^{1/a}$. For all rectangle $\mathcal{C}' \subset (0, +\infty)^2$ we define $\underline{R}_{\mathcal{C}',1} = \underline{R}_{\mathcal{C}',2}$.

8.4. Example 4. As in the preceding example, we consider $\xi > 0$, $a, b > 0$ and the rectangle $\mathcal{C} = [a_1, a_2] \times [b_1, b_2] \subset (0, +\infty)^2$. Our aim is to find a rectangle $\mathcal{R}$ containing (a, b) such that

$$\mathcal{C} \cap \mathcal{R} \subset \left\{(a', b') \in (0, +\infty)^2,\, h^2(f_{(a,b)}, f_{(a',b')}) \leq \xi\right\}.$$

For all positive numbers a', b',

$$h^2\left(f_{(a,b)}, f_{(a',b')}\right) \leq 2h^2\left(f_{(a,b)}, f_{(a,b')}\right) + 2h^2\left(f_{(a,b')}, f_{(a',b')}\right).$$

We derive from Theorem 7.6 of Ibragimov and Has'minskii (1981) that

$$h^2\left(f_{(a,b)}, f_{(a,b')}\right) \leq \frac{(b'-b)^2}{8} \sup_{t \in [\min(b,b'), \max(b,b')]} |\psi(t) - \psi(a+t)|$$

where ψ is defined in the preceding example. By using the monotony of the function $t \mapsto \psi(t) - \psi(a+t)$, we deduce that if $b' > b_1$,

$$h^2\left(f_{(a,b)}, f_{(a,b')}\right) \leq \begin{cases} 1/8\left(\psi(b) - \psi(a+b)\right)(b'-b)^2 & \text{if } b' \geq b \\ 1/8\left(\psi(b_1) - \psi(a+b_1)\right)(b'-b)^2 & \text{if } b' < b. \end{cases}$$

Similarly,

$$h^2\left(f_{(a,b')}, f_{(a',b')}\right) \leq \frac{(a'-a)^2}{8} \sup_{t\in[\min(a,a'),\max(a,a')]} \left|\psi\left(t\right) - \psi\left(b'+t\right)\right|.$$

Hence, if $a' \in [a_1, a_2]$ and $b' \in [b_1, b_2]$,

$$h^2\left(f_{(a,b')}, f_{(a',b')}\right) \leq \begin{cases} 1/8\left(\psi(a) - \psi(a+b_2)\right)(a'-a)^2 & \text{if } a' \geq a \\ 1/8\left(\psi(a_1) - \psi(a_1+b_2)\right)(a'-a)^2 & \text{if } a' < a. \end{cases}$$

We deduce from the above inequalities that we can take

$$\begin{aligned} \mathcal{R} &= \left[a - \sqrt{\frac{2\xi}{\psi(a_1)-\psi(a_1+b_2)}}, a + \sqrt{\frac{2\xi}{\psi(a)-\psi(a+b_2)}}\right] \\ &\quad \times \left[b - \sqrt{\frac{2\xi}{\psi(b_1)-\psi(a+b_1)}}, b + \sqrt{\frac{2\xi}{\psi(b)-\psi(a+b)}}\right]. \end{aligned}$$

As in the two last examples, we take $\underline{R}_{\mathcal{C}',1} = \underline{R}_{\mathcal{C}',2}$ for all rectangle $\mathcal{C}' \subset (0,+\infty)^2$.

8.5. Example 5. For all $m, m' \in \mathbb{R}$, $\lambda, \lambda' > 0$,

$$h^2\left(f_{(m,\lambda)}, f_{(m',\lambda')}\right) = \begin{cases} 1 - \frac{2\sqrt{\lambda\lambda'}}{\lambda+\lambda'} e^{-\frac{\lambda}{2}|m'-m|} & \text{if } m' \geq m \\ 1 - \frac{2\sqrt{\lambda\lambda'}}{\lambda+\lambda'} e^{-\frac{\lambda'}{2}|m'-m|} & \text{if } m' \leq m. \end{cases}$$

We consider $\xi > 0$ and aim at finding $\mathcal{R}$ containing (m, λ) such that

$$\mathcal{R} \subset \left\{(m', \lambda') \in \mathbb{R} \times (0,+\infty),\ h^2(f_{(m,\lambda)}, f_{(m',\lambda')}) \leq \xi\right\}.$$

Notice that if $m' \geq m$ and if

$$\frac{2\sqrt{\lambda\lambda'}}{\lambda+\lambda'} \geq \sqrt{1-\xi} \quad \text{and} \quad e^{-\frac{\lambda}{2}(m'-m)} \geq \sqrt{1-\xi}$$

then $h^2(f_{(m,\lambda)}, f_{(m',\lambda')}) \leq \xi$. Similarly, if $m' \leq m$ and if

$$\frac{2\sqrt{\lambda\lambda'}}{\lambda+\lambda'} \geq \sqrt{1-\xi} \quad \text{and} \quad e^{-\frac{\lambda'}{2}(m-m')} \geq \sqrt{1-\xi}$$

then $h^2(f_{(m,\lambda)}, f_{(m',\lambda')}) \leq \xi$. We can then take

$$\mathcal{R} = \left[m - \frac{1-\xi}{1+\xi+2\sqrt{\xi}} \frac{\log\left(1/(1-\xi)\right)}{\lambda}, m + \frac{\log\left(1/(1-\xi)\right)}{\lambda}\right] \times \left[\frac{1+\xi-2\sqrt{\xi}}{1-\xi}\lambda, \frac{1+\xi+2\sqrt{\xi}}{1-\xi}\lambda\right].$$

Let now $\mathcal{C}' = [\underline{m}_0, \bar{m}_0] \times \left[\underline{\lambda}_0, \bar{\lambda}_0\right]$ be a rectangle of $\mathbb{R} \times (0,+\infty)$. By proceeding as in the Gaussian model, we can define $\underline{\boldsymbol{R}}_{\mathcal{C}'}$ by

$$\underline{\boldsymbol{R}}_{\mathcal{C}'} = \left(\underline{R}_{\mathcal{C}',1}, \underline{R}_{\mathcal{C}',2}\right) = \left(\frac{2\sqrt{\bar{\lambda}_0\underline{\lambda}_0}}{\bar{\lambda}_0+\underline{\lambda}_0} \frac{1 - e^{-\underline{\lambda}_0(\bar{m}_0-\underline{m}_0)/2}}{\bar{m}_0 - \underline{m}_0}, \frac{1}{8\bar{\lambda}_0^2}\right).$$

8.6. Example 6. For all $m, m' \in \mathbb{R}$, $r, r' > 0$,

$$h^2\left(f_{(m,r)}, f_{(m',r')}\right) = 1 - \frac{(\min\{m+r, m'+r'\} - \max\{m, m'\})_+}{\sqrt{rr'}}$$

where $(\cdot)_+$ is the positive part of $(\cdot)$. We consider $\xi \in (0, \bar{\kappa})$, and aim at finding a rectangle $\mathcal{R}$ containing (m, r) such that

$$\mathcal{R} \subset \left\{(m', r') \in (0, +\infty)^2,\, h^2(f_{(m,r)}, f_{(m',r')}) \leq \xi\right\}.$$

For this, we begin to assume that $m' \leq m + r$ and $m' + r' \geq m$ to ensure that

$$h^2\left(f_{(m,r)}, f_{(m',r')}\right) = 1 - \frac{\min\{m+r, m'+r'\} - \max\{m, m'\}}{\sqrt{rr'}}.$$

Several cases are involved

- If $m' \leq m$ and $m' + r' \geq m + r$, a sufficient condition for $h^2\left(f_{(m,r)}, f_{(m',r')}\right) \leq \xi$ is

$$r' \leq \frac{1}{(1-\xi)^2} r.$$

- If $m' \geq m$ and $m' + r' \leq m + r$, a sufficient condition is

$$r' \geq (1-\xi)^2 r.$$

- If $m' \leq m$ and if $m' + r' \leq m + r$, a sufficient condition is

$$m - m' \leq \left(\sqrt{r'} - (1-\xi)\sqrt{r}\right)\sqrt{r'},$$

  which holds when

$$r' \geq (1-\xi/2)^2 r \quad \text{and} \quad |m' - m| \leq \xi/2\sqrt{1-\xi/2}\, r.$$

- If $m' \geq m$ and if $m' + r' \geq m + r$, a sufficient condition is

$$m' - m \leq \left(\sqrt{r} - (1-\xi)\sqrt{r'}\right)\sqrt{r}.$$

  This condition is fulfilled when

$$r' \leq \frac{1}{(1-\xi/2)^2} r \quad \text{and} \quad |m' - m| \leq \frac{\xi}{2-\xi} r.$$

We can verify that if (m', r') belongs to the rectangle

$$\mathcal{R} = \left[m - \frac{\xi\sqrt{2-\xi}}{2\sqrt{2}} r, m + \frac{\xi}{2-\xi} r\right] \times \left[(1-\xi/2)^2 r, \frac{r}{(1-\xi/2)^2}\right]$$

then $m' \leq m + r$ and $m' + r' \geq m$ (since $\xi \leq \bar{\kappa}$). The rectangle $\mathcal{R}$ suits. For all rectangle $\mathcal{C}' \subset (0, +\infty)^2$ we choose in this example $\underline{R}_{\mathcal{C}',1} = \underline{R}_{\mathcal{C}',2}$.

8.7. Speed of the procedure. By way of indication, we give below the number of tests that have been calculated in the simulation study of Section 6.

	$n = 25$	$n = 50$	$n = 75$	$n = 100$
Example 1	1602 (117)	1577 (72)	1570 (60)	1567 (52)
Example 2	2935 (90)	2937 (76)	2938 (69)	2938 (64)
Example 3	9082 (1846)	8700 (1183)	8569 (934)	8511 (800)
Example 4	10411 (778)	10272 (461)	10236 (357)	10222 (304)
Example 5	6691 (296)	6699 (210)	6715 (175)	6726 (158)
Example 6	32614 (1238)	33949 (1211)	34822 (1190)	35397 (1177)

Figure 4.7: Number of tests computed averaged over 10^4 samples and their corresponding standard deviations in brackets.

References

Anatolyev, S. and Kosenok, G. (2005). An alternative to maximum likelihood based on spacings. *Econometric Theory*, 21(2):472–476.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401.

Baraud, Y. (2013). Estimation of the density of a determinantal process. *Confluentes Mathematici*, 5(1):3–21.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284.

Barnett, V. D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, 53(1/2):151–165.

Basu, A. K., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Basu, A. K. and Lindsay, B. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Probability Theory and Related Fields*, 65:181–237.

Birgé, L. (1984a). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, 20:201–223.

Birgé, L. (1984b). Sur un théorème de minimax et son application aux tests. *Probability and Mathematical Statistics*, 2:259–282.

Birgé, L. (2004). Model selection for Gaussian regression with random design. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 10(6):1039–1051.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2012). Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, volume 9, pages 47–64. IMS Collections.

Birgé, L. (2013). Model selection for density estimation with $\mathbb{L}_2$-loss. *Probability Theory and Related Fields*, pages 1–42.

Cheng, R. and Amin, N. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):394–403.

Dacunha-Castelle, D. (1978). Vitesse de convergence pour certains problèmes statistiques. In *École d'Été de Probabilités de Saint-Flour, VII (Saint-Flour, 1977)*, volume 678 of *Lecture Notes in Math.*, pages 1–172. Springer, Berlin.

Ekström, M. (1998). On the consistency of the maximum spacing method. *Journal of Statistical Planning and Inference*, 70(2):209–224.

Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, 77(380):831–834.

Ghost, K. and Jammalamadaka, S. R. (2001). A general estimation method using spacings. *Journal of Statistical Planning and Inference*, 93(1):71–82.

Ibragimov, I. and Has'minskii, R. (1981). *Statistical estimation–asymptotic theory*. Applications of mathematics. Springer-Verlag.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54. Academic Press, New York.

Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review*, pages 153–171.

Lindsay, B. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22(2):1081–1114.

Pitman, E. (1979). *Some basic theory for statistical inference*. Monographs on applied probability and statistics. Chapman and Hall.

Ranneby, B. (1984). The maximum spacing method. an estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, 11(2):93–112.

Shao, Y. and Hahn, M. G. (1999). Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Annals of the Institute of Statistical Mathematics*, 51(1):31–49.

Estimation par tests

Résumé. Cette thèse porte sur l'estimation de fonctions à l'aide de tests dans trois cadres statistiques différents. Nous commençons par étudier le problème de l'estimation des intensités de processus de Poisson avec covariables. Nous démontrons un théorème général de sélection de modèles et en déduisons des bornes de risque non-asymptotiques sous des hypothèses variées sur la fonction à estimer. Nous estimons ensuite la densité de transition d'une chaîne de Markov homogène et proposons pour cela deux procédures. La première, basée sur la sélection d'estimateurs constants par morceaux, permet d'établir une inégalité de type oracle sous des hypothèses minimales sur la chaîne de Markov. Nous en déduisons des vitesses de convergence uniformes sur des boules d'espaces de Besov inhomogènes et montrons que l'estimateur est adaptatif par rapport à la régularité de la densité de transition. La performance de l'estimateur est aussi évalué en pratique grâce à des simulations numériques. La seconde procédure peut difficilement être implémenté en pratique mais permet d'obtenir un résultat général de sélection de modèles et d'en déduire des vitesses de convergence sous des hypothèses plus générales sur la densité de transition. Finalement, nous proposons un nouvel estimateur paramétrique d'une densité. Son risque est contrôlé sous des hypothèses pour lesquelles la méthode du maximum de vraisemblance peut ne pas fonctionner. Les simulations montrent que ces deux estimateurs sont très proches lorsque le modèle est vrai et suffisamment régulier. Il est cependant robuste, contrairement à l'estimateur du maximum de vraisemblance.

Mots clés : Estimation paramétrique, Chaîne de Markov, Sélection de modèles, Statistiques non-asymptotiques, Processus de Poisson, Robustesse, Sélection d'estimateurs, T-estimateur.

Estimation via testing

Abstract. This thesis deals with the estimation of functions from tests in three statistical settings. We begin by studying the problem of estimating the intensities of Poisson processes with covariates. We prove a general model selection theorem from which we derive non-asymptotic risk bounds under various assumptions on the target function. We then propose two procedures to estimate the transition density of an homogeneous Markov chain. The first one selects an estimator among a collection of piecewise constant estimators. The selected estimator is shown to satisfy an oracle-type inequality under minimal assumptions on the Markov chain which allows us to deduce uniform rates of convergence over balls of inhomogeneous Besov spaces. Besides, the estimator is adaptive with respect to the smoothness of the transition density. We also evaluate the performance of the estimator in practice by carrying out numerical simulations. The second procedure is only of theoretical interest but yields a general model selection theorem from which we derive rates of convergence under more general assumptions on the transition density. Finally, we propose a new parametric estimator of a density. We upper-bound its risk under assumptions for which the maximum likelihood method may not work. The simulations show that these two estimators are very close when the model is true and regular enough. However, contrary to the maximum likelihood estimator, this estimator is robust.

Keywords: Estimator selection, Markov chain, Model selection, Non-asymptotic statistics, Parametric estimation, Poisson processes, Robustness, T-estimator.